

# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



**MAGNETIC  
SOFT MATERIALS  
TRANSFORM THEIR  
SHAPE IN FRACTIONS  
OF A SECOND** **PAGE 274**

## QUICK-CHANGE ARTISTS

GLOBAL HEALTH

### PANDEMIC PREVENTION

*Human surveillance will  
trump viral genomics*

**PAGE 180**

QUANTUM PHYSICS

### GOING THE DISTANCE

*Taking the element of chance  
out of quantum networks*

**PAGES 192, 264 & 268**

BIOPHYSICS

### CATALYSIS IN A COLD CLIMATE

*Secrets of enzyme function  
at low temperatures*

**PAGES 195 & 324**

**NATURE.COM/NATURE**

14 June 2018

Vol. 558, No. 7709



# THIS WEEK

## EDITORIALS

**LANGUAGE** The illogical misuse of 'necessary and sufficient' **p.162**

**WORLD VIEW** It makes sense for Antarctic scientists to share **p.163**



**MATERIALS** Solar cells from seashells on the seashore **p.165**

## Reform the Antarctic Treaty

*Political protection for the planet's last great wilderness is no longer fit for purpose. Make its governance democratic: scrap the veto that lets individual interests rule.*

Of the common adjectives used to describe Earth's southern polar region, 'pristine' is among the most inappropriate. The ocean around Antarctica bobs with pieces of microplastic pollution, and for decades, whales and other marine life have been stripped from the sea. The ozone hole gapes above. To find any of the advertised unspoiled wilderness, a visitor has to trek inshore and away from the direct influence of the rest of the world. Because there is another misapplied label: remote. It might look isolated on a map, but the Antarctic is within increasingly convenient reach — for good and for bad.

Campaigners last week said that Antarctic snow samples they had gathered were polluted with persistent hazardous chemicals. And figures presented at the annual meeting of parties to the Antarctic Treaty last month in Buenos Aires showed that 45 private yachts were spotted in sensitive Antarctic waters — or reported an intention to visit — over the most recent southern summer season. That's up by one-third on the previous year. Nine did so without permission, and crew and passengers on at least one were seen to violate strict protections by approaching birds' nests, flying drones through rookeries and touching animals.

The tasks of setting rules to control all this behaviour and working to protect the continent from (further) harm fall heavily on delegates from dozens of countries who attend those Antarctic Treaty meetings. And the Buenos Aires gathering did have some success, drafting new rules on drone use and settling other minor issues. But when it comes to measures to address the bigger challenges, not least how to conserve fish and the other marine life that survives in the Southern Ocean, the treaty is at the mercy of geopolitics — and there are worrying signs that it is struggling to cope.

*Nature* this week publishes a series of articles discussing this and other issues that are emerging in Antarctica. An Insight supplement explores the mechanisms that control the movement of Antarctica's ice and interactions with the broader climate system (see page 199). A Comment describes the perilous position of Antarctic fisheries (see page 177). And a World View (see page 163) makes the case for stronger collaboration and sharing of research infrastructure between scientists from different countries who work in Antarctica.

Pressure on the Antarctic Treaty from geopolitics can only increase, as demand for the continent's stocks of fish and expected reserves of minerals rises with the depletion of resources elsewhere. Formally, the treaty prohibits mining until at least 2048. But the protection that the agreement offers increasingly relies on good will as much as international law, as shown by the determination of some nations to continue fishing in Antarctic waters, in the face of proposals to ban the practice by establishing reserves and protected zones.

In one sense, it's difficult for science to lament the unwelcome intrusion of international politics into its Antarctic playground. The 1959 creation of the Antarctic Treaty — which fenced the territory off for research — was itself an act of supreme realpolitik, a way for a resurgent post-war United States to gain influence on a continent it had

previously ignored. In agreeing to set aside their overlapping territorial claims — for example, the United Kingdom, Chile and Argentina all contest the same region — the other founding nations also found the treaty convenient, because they could put off any attempts to resolve their disputes without losing face. It's partly for this reason that the treaty is suggested as a model of possible future governance in regions of tension elsewhere, such as the South China Sea.

But although science has been the face of national expansion into the Antarctic in the decades since the treaty began, strategic interests

***"The treaty is at the mercy of geopolitics — and there are worrying signs that it is struggling to cope."***

have always been close behind. And now that they are threatening to take the lead, what can science do to protect its own needs in the region — as well as the planet's last remaining chunk of relatively unspoiled land, in the Antarctic interior?

First, the scientific community should recognize the scale of the challenge. The Antarctic Treaty has stood for a long time and although it might look solid, it is fragile and vulnerable to special interests — a bit like Antarctica itself. Some 53 nations now contribute to the governance of Antarctica through the treaty system, and not as a democracy. Individual countries can veto measures they dislike, allowing them to continue activities that the majority wish to outlaw, which is one reason why the system has not produced any new binding protocols (measures to enforce the treaty's principles) for two decades. Just as the original countries agreed to postpone arguments over their national claims when they signed the treaty, current members prefer to kick difficult decisions down the road. But it's often those decisions — such as how to punish nations that break the rules — that scientists and other Antarctic advocates care about the most.

That's not to say that science has lost its voice. If anything, the opposite is true. An entire generation across the world has grown up seeing science as a priority for Antarctica under the treaty. And most of those people would surely object to the idea of science and conservation being tossed aside so that the Antarctic wilderness can be fished, mined, polluted or developed. Scientists can strengthen and harness such support by relentlessly telling the public and policymakers about the seriousness of the threat to Antarctica and the need to protect the region. The stakes are high: this really is the last chance we have to leave a piece of the planet close to the way we found it.

Ultimately, changes in governance will probably be required to maintain the primacy of science over exploitation in Antarctica, and to minimize the environmental damage that changes to the polar region will cause to the rest of the planet. The Antarctic Treaty was a triumph of global politics, but global politics has changed. Even though any voting system is subject to gaming, the time of the single-country veto has passed. Let the future of Antarctica be decided by the majority. ■



# Chinese checkers

*China sets a strong example on how to address scientific fraud.*

The Chinese government knows that a slice of its generous science budget — the world's second-largest by country — goes to waste on bad science. It doesn't want to waste any more. On 30 May, the State Council and the Communist Party of China announced a radical new system of regulations to police science and raise research standards in the country.

Certainly, reform is necessary and overdue. Various Chinese government bodies have made the case to crack down on fraud and misconduct in science over the past two decades, but with limited success. This time, the changes have serious political weight behind them and could make a significant difference. The policy might offer the greatest disincentive to cheating in research that the world has seen so far. But the devil, as always, will be in the detail — and in how well the plans are enforced.

One of the most striking conditions is that researchers will be deterred from publishing findings in journals that China deems to be of poor academic quality, poorly managed and set up merely for profit. Many such 'predatory journals' offer researchers a place to publish, for a fee, and shirk their editorial responsibility to evaluate papers to determine quality. China's science ministry is working on a blacklist of those journals. In an unprecedented step, any researcher who publishes in one will get a warning and be given no credit for the publication when they are evaluated for grants or jobs. Using government grants to pay the publication fees in these journals, as many presumably do, could land Chinese scientists in deeper trouble.

As the world's largest producer of scientific papers, China's new rules could go as far as to put some of these rogue journals out of business, and that could be good for scientists everywhere. (Although, as we discuss in a News story this week on page 171, some scientists are anxious about how these journals are identified, while others have concerns

about such blacklists and prefer 'whitelists' of approved publications.)

In another major shift, China is handing the responsibility for deterring and investigating scientific misconduct to the government's science ministry. That's quite a shake-up for China, where — as in many places — institutions are usually expected to investigate allegations against their own researchers. That is too often ineffective. With little to gain and a reputation to lose, many prefer to sit on their hands and wait for the situation to blow over.

**"The policy might offer the greatest disincentive to cheating in research that the world has seen."**

Denmark, for instance, has designated a national agency to police science, but, too frequently, there is limited will and scant resources to pursue allegations of fraud at the government level. In the United States, for example, the Office of Research Integrity is short-staffed and has limited leverage over universities.

In China, the situation could play out differently. The new rules state explicitly that institutions that shield errant scientists can be punished through a loss of funding. That could give the policy real teeth — enough to drastically clean up Chinese research. But success will take sustained effort and pressure from the top, and because there is no guarantee of that, the policy could equally fall flat. China's bureaucrats are not responsive to its citizens — no matter how loud the cry on social media for an investigation into a given scientist — and they make almost no effort to be transparent. The science ministry could stick its head in the sand just as deeply as some institutions do.

There are other causes for concern. The science ministry is also drawing up rules on how penalties will be meted out — including the blacklisting of scientists who have committed particularly egregious acts. To maintain fairness, harsh penalties require assurances that the judgements leading to them are based on thorough and fair evaluations.

China's bureaucrats might not answer to the people, but they do answer to the higher echelons of power. The current push for better management of science comes as part of President Xi Jinping's wider anti-corruption drive. Xi regularly talks up the crucial role of science and technology in making China stronger and more independent. With its new rules, China is backing words with actions. ■

# Not necessary

*Phrase 'necessary and sufficient' blamed for flawed neuroscience.*

In his 1946 classic essay 'Politics and the English language', George Orwell argued that "if thought corrupts language, language can also corrupt thought". Can the same be said for science — that the misuse and misapplication of language could corrupt research? Two neuroscientists believe that it can. In an intriguing paper published in the *Journal of Neurogenetics*, the duo claims that muddled phrasing in biology leads to muddled thought and, worse, flawed conclusions (M. Yoshihara and M. Yoshihara *J. Neurogenet.* **32**, 53–64; 2018).

The phrase in the crosshairs is "necessary and sufficient". It's a popular one: figures suggest the wording pops up in some 3,500 scientific papers each year across genetics, cell biology and neuroscience alone. It's not a new fad: *Nature's* archives show consistent use since the nineteenth century.

Used properly, the phrase indicates a specific relationship between two events. For example, the statement, "I'll pay for lunch if, and only if, you pay for breakfast," can be written as, "You paying for breakfast is necessary and sufficient for me paying for lunch."

But, argue Motojiro Yoshihara and Motoyuki Yoshihara, use of the phrase in research reports is problematic, and should be curtailed.

The logic of the term is at the heart of the dispute. It's too often used

as shorthand to mean 'linked to' or 'important for', the authors say. And this sloppy use, they argue, can lead scientists in the wrong direction, especially in genetics.

If a gene is necessary and sufficient for something (as often claimed), strict logic demands that that gene alone can do the job. For example, the gene *eyeless* is certainly necessary for a retina to develop. But it is not sufficient — if it were, then logic would demand that 'if *eyeless* exists, then a retina will develop'. This is false; other genes and factors are needed as well. Yet *eyeless* is often described as being necessary and sufficient for retinal development.

The duo argues that its objection to such incorrect use is more than pedantry. The combination of necessary and sufficient is excessively strict, and its widespread use has meant, for example, that some 'command' neurons have failed to be identified as such because they don't satisfy the required criteria. (The agreed definition of a command neuron is one that is necessary and sufficient to initiate a behaviour.)

One such missed neuron is the Mauthner cell, responsible for a fast-escape reflex in fishes and amphibians. In fact, so few command neurons satisfy the logic of the phrase that the concept that they exist at all has been undermined, the authors say.

In most cases, they propose, a better phrase than 'necessary and sufficient' would be "indispensable and inducing". (Number of uses so far: one, in their paper.)

Will it catch on? Biologists will no doubt counter that they use the 'necessary and sufficient' phrase in a mutually understood way that is separate from its logical roots. Perhaps, but then Orwell had that covered, too: "A bad usage can spread by tradition and imitation even among people who should and do know better." ■



SCAR



## Polar collaborations are key to successful policies

*Expand the remit of the Scientific Committee on Antarctic Research to coordinate the influx of infrastructure, says its president, Steven L. Chown.*

The Antarctic Treaty, now signed by 53 countries, enshrines the idea of a continent free of discord and set aside for science. In 1959, nations pledged to give up territorial claims on the region, to use the continent only for peaceful purposes, to cooperate on scientific investigations and to freely share results. An addendum called the Environmental Protocol to protect the region and prohibit mining came into force in 1998. No other major agreements under the treaty have been enforced since (see page 161).

Meanwhile, research infrastructure is booming. Nations including China, Germany, South Korea, New Zealand and the United Kingdom have been building new research stations or upgrading existing ones. Australia, Norway and the United Kingdom have launched or will soon launch research vessels much larger than their predecessors. In May, Australia announced that it would, subject to environmental approvals, build the first paved runway on the continent, a 2.7-kilometre airstrip that will open up East Antarctica to large aircraft all year round. Several other countries are also upgrading their air access.

Why all the interest? Understanding the region is essential to predicting global environmental change, particularly sea-level rise — with its implications for people and property. And there are geopolitical considerations, too. More countries want to maintain a presence below the 60th parallel south, especially as the Southern Ocean becomes more accessible to fishing (see page 177).

But too few nations are building the intellectual capacity to complement the accumulating infrastructure. They should be looking for ways to share research resources and coordinate efforts. Diffuse activity will not provide the kind of information that world leaders can act on. It is time to focus.

The Scientific Committee on Antarctic Research (SCAR), which I head, was established almost at the same time as the treaty, and acts as a scientific adviser for the region. It already coordinates activities ranging from marine-predator surveys to space-weather forecasts, which provide geomagnetic-storm warnings that are necessary to secure electricity grids and satellites.

SCAR is often asked for advice on everything from bioprospecting to the potential influence of drones on wildlife, but is rarely funded to provide it. Countries that readily invest millions of dollars in infrastructure often struggle to find funds for policy-relevant science. They would get a much better return on their investment if they put a tiny fraction of infrastructure spending towards coordinated research.

The remit of SCAR should be formally expanded to coordinate resources more broadly and to formulate scientific questions whose answers could shape global agreements. One crucial question is how long the Southern Ocean will continue to take up carbon as its waters warm and acidify. Another is how changing krill populations will alter ecosystems, which include iconic predators such as penguins,

seals and whales, and threaten fishing grounds.

The most urgent task is to understand the Antarctic ice sheets well enough to reduce uncertainties about sea-level change. More research is needed into how ice shelves buttress ice sheets, how the ocean, ice shelf and atmosphere interact, how melting water fractures ice shelves and how snowfall on the continent is changing.

Clarity on these fronts is needed to hold signatories of the Paris climate agreement to their promises to reduce greenhouse-gas emissions and to help adapt to coming changes. Without good models, much planning effort will be wasted. In fact, without better information from the Antarctic, global human populations near the coasts will not be able to work out how to accommodate higher sea levels and more-frequent storm surges. Failure to plan appropriately will make it harder to act not just on climate change but also on migration, justice and conservation.

Gaining the needed clarity will require shared infrastructure and coordination. This could take the form of joint research cruises to collect sediment and ice cores, sub-ice-shelf investigations by autonomous vehicles and modelling to improve estimates of changes in local ice-shelf and ice-sheet behaviour across the Antarctic.

In April, the United Kingdom and United States announced they would work together to study the broad, fast-moving Thwaites Glacier flowing into the Amundsen Sea. Other researchers could join up throughout the region, and SCAR could facilitate and support more such efforts, align modelling and empirical data gathering and ensure that data and infrastructure, such as ships and stations, are readily available.

SCAR has a track record of facilitating successful collaborations. These include the discovery of the ozone-layer hole and elucidation of its chemistry, and clarification of the relative importance of food-web paths in Southern Ocean ecosystems. Doing such science is, however, quite different from ensuring that its results inform policy.

Happily, SCAR has another unique capability, one that drew me to the organization. The committee is a designated adviser to the parties to the Antarctic Treaty, and their delegations are well placed to influence decisions about the region and its neighbours. SCAR can also advise global bodies such as the Intergovernmental Panel on Climate Change, the United Nations Framework Convention on Climate Change and the Convention on Biological Diversity.

In other words, SCAR is poised to coordinate research effectively and to convey its findings to the bodies that are best able to act on the information. Parties to the treaty should seize this unique opportunity to support science that can bring better decisions for the planet. ■

**AUSTRALIA  
ANNOUNCED IT  
WOULD BUILD THE  
FIRST PAVED  
RUNWAY  
ON THE CONTINENT.**

Steven L. Chown is professor of biological sciences at Monash University in Melbourne, Australia, and president of SCAR.  
e-mail: [steven.chown@monash.edu](mailto:steven.chown@monash.edu)



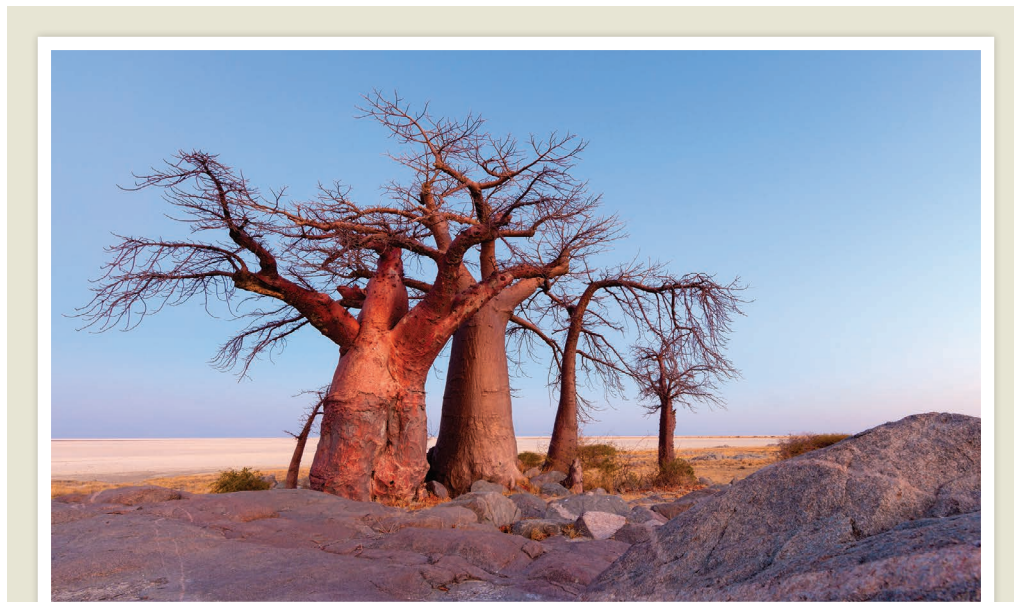
## PEOPLE

### Astronaut minister

Astronaut Pedro Duque was appointed Spain's minister of science, innovation and universities on 7 June. The country's socialist party formed a new minority government after a vote of no confidence ousted the previous prime minister. Spain's government science budget plunged by more than 30% after the 2008 financial crisis, and scientists say that austerity measures have added bureaucracy around research expenses. The government has committed to sticking with an austerity budget for 2018, but Duque says that he will try to ease some cost-controlling rules that delay laboratory purchases and the spending of grants. Duque was Spain's first astronaut, with the European Space Agency. He flew on the space shuttle *Discovery* in 1998 and to the International Space Station in 2003, where he ran science experiments.

### US science envoys

A former astronaut is among five science envoys announced by the US Department of State on 8 June. In a programme begun under former US president Barack Obama, the envoys travel the world to strengthen international cooperation in science and technology. The new class comprises former NASA chief Charles Bolden, who flew aboard four space-shuttle missions; Robert Langer, a chemical engineer at the Massachusetts Institute of Technology in Cambridge; Michael Osterholm, an infectious-disease researcher at the University of Minnesota in Minneapolis; Rebecca Richards-Kortum, a bioengineer at Rice University in Houston, Texas; and James Schauer, who studies air quality at the University of



HOUGAARD MALAN/NATUREPL.COM

## Africa's iconic baobab trees are dying

The oldest and largest baobab trees in Africa are dying — but no one knows why. The iconic African baobab tree (*Adansonia digitata*), which sports a wide trunk and high branches, is Earth's oldest living flowering plant, and it can live for 2,000 years. In a study intended to investigate the age, structure and longevity of Africa's largest and potentially oldest baobabs, researchers were surprised to find that several of

the trees died during the study period. Between 2005 and 2017, they dated more than 60 trees and found that during that period, 9 of the oldest 13 trees died, as did 5 of the 6 largest ones. The researchers, who published their findings on 11 June, found no signs of an epidemic or disease; they suggest that changes in climate could be to blame (A. Patrut *et al. Nature Plants* <http://doi.org/cqsh>; 2018).

Wisconsin–Madison. Such appointments are rare for the administration of current US President Donald Trump, which has appointed very few science advisers.

### Verma resigns

Cancer researcher Inder Verma has resigned from the Salk Institute for Biological Sciences in La Jolla, California, the institute said on 11 June. Verma had been temporarily suspended since 21 April, after several female scientists with ties to the institute alleged that he had sexually harassed and discriminated against them. The Salk Institute then began an investigation of the complaints, which Verma has repeatedly denied, and on

11 June its board of trustees met to discuss the findings of that probe. According to the institute, Verma resigned before the Salk board of trustees had concluded its investigation. At press time, Verma had not responded to *Nature's* request for comment.

## FUNDING

### Private money

US research institutions received more than US\$2.3 billion for basic science from philanthropic organizations and corporations in 2017, according to a survey by the Science Philanthropy Alliance in Palo Alto, California. Now in its third year, the survey

looked at 46 institutions and showed that private funding for basic science remained flat from 2016 to 2017. Among the 25 institutions for which the alliance has data from all three years, however, funding rose to \$1.7 billion in 2017, a 13% increase from 2016 and a 40% increase from the \$1.2 billion that the survey registered in 2015.

### Hawking fellows

The United Kingdom has established a postdoctoral fellowship scheme in honour of the late physicist Stephen Hawking. Over the next five years, national science funder UK Research and Innovation will award up to ten grants each year for



JOHN THYS/AP/GETTY exceptional graduate students in mathematics, physics and computer science. The funding, details of which are yet to be released, will allow the students to continue the work from their doctorates at any institution in the country for up to three years. Science minister Sam Gyimah said that the fellowships offer a “fitting tribute” to Hawking, who died in March and whose work changed our understanding of the Universe.

## RESEARCH

## Record flooding

High-tide flooding in US coastal areas is twice as frequent as it was 30 years ago, according to a report released by the National Oceanic and Atmospheric Administration on 6 June. Tide gauges at 98 locations around the country (excluding Alaska) measured a record-high average of 6 flood days between May 2017 and April 2018. Sabine Pass, Texas, recorded the highest number of flood days, at 23. Although some of the flooding coincided with storms, much of it was a result of sea-level rise.

## POLICY

## European science

The European Commission has outlined how it plans to spend the biggest research and innovation budget in its history. On 7 June,



Carlos Moedas, the European Union's commissioner for research (pictured), presented a detailed proposal for the structure of the Horizon Europe science-funding programme, which will run from 2021 to 2027 and has a provisional budget of nearly €100 billion (US\$117 billion). Horizon Europe will have three main pillars: one each serving basic research and innovation, and another directed at solving societal challenges and at boosting industrial competitiveness. The programme will be open to all countries worldwide for the first time, giving the United Kingdom an opportunity to take part after Brexit. See [go.nature.com/2llw3jf](http://go.nature.com/2llw3jf) for more.

## Wild horses spared

Science and environment groups in Australia have slammed new laws that will protect wild horses in New South Wales' largest national park. State politicians passed legislation to protect the

heritage value of the horse — an invasive species known as a brumby — on 6 June. It rules out a cull recommended by the state government's own environment department to reduce the number of horses to preserve delicate ecosystems in the Kosciuszko National Park. The Australian Academy of Science says that the laws go against scientific advice, while the International Union for Conservation of Nature warns that the bill sets a disturbing precedent. Its passing prompted ecologist David Watson — a member of the state government's scientific advisory committee on threatened species — to resign in protest.

## EVENTS

## Biotech start-up

A non-profit institute launched by the Bill & Melinda Gates Foundation earlier this year got its official introduction in a 7 June announcement at the BIO convention in Boston, Massachusetts. The Bill & Melinda Gates Medical Research Institute, which will have offices in both Boston and Seattle, Washington, will focus on diseases of poverty in low- and middle-income countries. With US\$273 million for its first four years, the group's researchers will probe findings from early-stage studies conducted at universities and biotechnology companies in

search of promising leads on cures for tuberculosis, malaria and diarrhoea — and aim to develop them until they are ready for clinical trials. The institute will then pass them to companies in return for commitments to make the products available to those in need.

## Malaria progress

Paraguay is malaria-free, the World Health Organization (WHO) said on 11 June. The South American country hasn't had a single case of malaria in the past six years. The WHO reported the news in an update on 21 countries that are aiming to eliminate the disease by 2020. The report highlights progress in China, El Salvador and Algeria, which each had fewer than ten cases last year. But South Africa, Botswana and several other countries hoping to eliminate the disease have instead seen a rise in cases. Resurgences are often caused by gaps in health care and lapses in mosquito control.

## SPACE

## Space missions

On 6 June, India's government gave its space agency approval to help Oman build a space programme. A memorandum of understanding, signed by the two nations in February, calls for cooperation in space science, planetary exploration, satellite-based navigation and remote sensing of Earth. Oman joins a growing number of countries developing space programmes. The neighbouring United Arab Emirates announced last week that it has decided on a shortlist for its astronaut-training initiative. The list of 95 men and women will be reduced by the end of the year to 4 astronauts, who will participate in international space missions from 2021. The country is also planning an uncrewed mission to Mars in 2020.

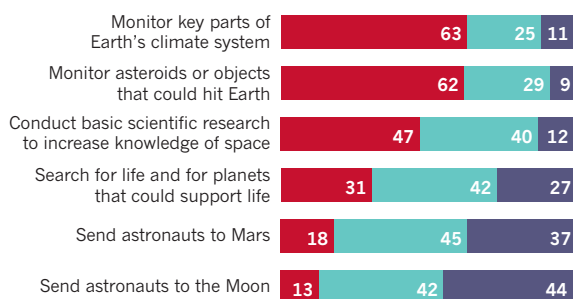
## TREND WATCH

About 70% of people in the United States think it is essential that their country remains a world leader in space exploration, a national survey of 2,541 people finds. And, in a time when the private space sector is growing, 65% say that NASA should still play a key part in space discovery. Respondents said that NASA's top priority should be monitoring Earth's climate system. But the idea of sending astronauts to the Moon or Mars — two priorities of Donald Trump's administration — didn't prove popular.

## WHAT SHOULD NASA DO?

In a survey of 2,541 US adults, most identified climate monitoring as a top priority for NASA.

■ Top priority ■ Important but lower priority ■ Not too important or should not be done



\*Percentages do not add up to 100% because of rounding.

Percentage of 2,541 respondents\*

# NEWS IN FOCUS

**CHINA** Sweeping reforms introduced to deal with scientific misconduct **p.171**

**DISEASE** Speedy Ebola tests help contain latest African outbreak **p.172**

**CONSERVATION** Can switching animals' gut bacteria make them more adaptable? **p.173**



**PHYSICS** Researchers debate how to assess discoveries of new elements **p.175**

GREG KENDALL-BALL/NATURE



Colombia has greater biodiversity than any country except Brazil.

## ECOLOGY

# Peace is killing Colombia's jungle — and opening it up

*Scientists are racing to document once-inaccessible regions as industry moves in.*

BY SARA REARDON

When the government of Colombia and the left-wing guerrilla group Revolutionary Armed Forces of Colombia (FARC) signed an agreement in 2016 to end five decades of conflict, the world celebrated. But that hard-won peace has come with a hidden cost.

FARC guerrillas once occupied large swathes of Colombia's vast forests, including the Colombian Amazon. The fighters'

presence sent smallholder farmers fleeing to cities and discouraged development. But as FARC has moved out of the forests, industry has moved in — including logging, gold-mining and cattle-grazing. A government analysis found that deforestation increased by 44% in the year of the peace accords.

Now, scientists are racing to document Colombia's rich biodiversity, which is second only to that of Brazil. In the process, they are rediscovering ecosystems that were once largely off-limits. Earlier this year, more than

40 researchers launched a digital platform to collate information on weather patterns, species distribution and other indicators of environmental health across Colombia. Their goal is to predict how the encroaching development could alter the country's forests and other ecosystems — information that could shape policies on land use, water security and other environmental issues as Colombia adjusts to peace.

If the researchers succeed, their efforts could yield insights that benefit other tropical ►





Plant samples collected by researchers at the Humboldt Institute in Bogotá.

nations grappling with climate change and development. “These topics are relevant not just for Colombia,” says Alejandro Salazar, a soil biologist at Purdue University in West Lafayette, Indiana.

For decades, researchers who wanted to study ecosystems in areas controlled by FARC and other armed groups had to get permission from the fighters. Some scientists persevered. Between 1988 and 2002, primatologist Pablo Stevenson ran a field station in the Macarena region, which is home to several primate species and was then under FARC control. The guerrillas would occasionally stop in for coffee, but generally left Colombian scientists alone, says Stevenson, now at the University of the Andes in Bogotá. “They were very respectful.”

Foreign researchers had a harder time — possibly because FARC thought they would command a high ransom. In 2002, Stevenson closed his field station, after guerrillas kidnapped a Japanese primatologist and demanded that Colombian universities pay to get him back. Still, the FARC guerrillas allowed the primatologist to keep working during his three-month captivity. One US biologist who was kidnapped twice by FARC even named a toad species *Atelopus farci*, because its mottled green skin resembled the guerrillas’ uniforms.

But many scientists avoided such field trips altogether. Now, the peace agreement has got them excited about the future. “We’ve been exposed to science in conference rooms,” says Daniel Ruiz Carrascal, a geoscientist at the Antioquia School of Engineering in Medellín. “We haven’t had the chance to go to the field.” He has begun building a network of

temperature and humidity sensors in alpine regions to monitor climate change.

The first wave of peacetime studies has already yielded surprising discoveries. Last year, biologists at the Alexander von Humboldt Biological Resources Research Institute in Bogotá found six new species, including frogs and beetles, after just ten days of searching in one forested area near Medellín.

To paint a broader picture of the environment, Ruiz Carrascal and others launched the Platform for Ecological Analyses on Colombian Ecosystems (PEACE). Their goal is to connect data from satellites, monitoring systems and the small number of scientists who conducted fieldwork during the FARC conflict.

The PEACE consortium’s main project is a ‘datacube’: an environmental model that simulates future change using information on atmospheric conditions, forest cover, fires and animal populations, among other factors. Researchers want to use the model to answer questions such as how ecosystems adapt to unusual conditions in years when the El Niño weather pattern occurs. The researchers also hope to reveal how increased deforestation and human migration caused by Colombia’s peace process might affect the environment. Such data could inform policies to protect ecologically important areas and limit exploitation of natural resources.

Other questions about the future of conservation efforts centre on Colombia’s dismal record of enforcing its environmental laws.

Eighty-four per cent of the deforestation there so far has taken place in protected areas, says Juan Posada, an ecologist at El Rosario University in Bogotá. The continued presence of illegal armed groups further complicates matters. “There are rules in place for biodiversity, but no one is respecting them,” he says.

And any conservation efforts also have to deal with the social issues that have arisen as Colombia tries to recover from decades of conflict. Since the early 1960s, when FARC emerged as a national force, more than 7 million people have fled their homes, largely in rural areas. Landowners who try to return to their abandoned plots are facing an unexpected problem. Jungles that were once cleared for farmland grew back during FARC occupation; the government has claimed much of this land as a natural resource that cannot be used for farming or timber. This has left hundreds of thousands of people with no source of income if they choose to move back to their land.

“This is a part of the conflict nobody saw coming,” says Carlos Zuluaga, director-general of the Antioquia branch of Colombia’s environmental protection agency, CORNARE.

In 2013, CORNARE launched a programme that pays about 3,000 families up to 600,000 Colombian pesos (US\$200) a month to preserve their land, rather than developing it. Zuluaga says that some of these areas could be worth more to Colombia if they are kept ecologically intact, because of their role in the water and carbon cycles.

Paying people to protect ecosystems has been tried in other countries, such as Brazil and Uganda. It can work well, at least in the short term, says Jennifer Alix-Garcia, an economist at Oregon State University in Corvallis. “The long-term sustainability is kind of unknown because most haven’t been in place long enough,” she says. The key, she says, is for governments to choose the areas to protect carefully, and to enforce their agreements with landowners. Otherwise, people could take the government’s money even as they give developers access to their land.

Even if enforcement remains weak, the PEACE researchers hope that their project — and any monitoring sites they set up — will at least help policymakers to understand the scope of the problem. “We are in a way rediscovering our country,” Ruiz Carrascal says. ■

Travel for this story was paid for by the Pulitzer Center on Crisis Reporting in Washington DC.

GREG KENDALL-BALL/NATURE

  
**MORE  
ONLINE**

#### TOP NEWS



Northern fur seals can go weeks without REM sleep  
[go.nature.com/2i68rbj](http://go.nature.com/2i68rbj)

#### MORE NEWS

- Honeybees can count to zero  
[go.nature.com/2jnzofh](http://go.nature.com/2jnzofh)
- EU to world: join our €100-billion research programme  
[go.nature.com/21lw3jf](http://go.nature.com/21lw3jf)
- US study: sexual harassment is rife in the sciences  
[go.nature.com/2jdyavy](http://go.nature.com/2jdyavy)

#### NATURE PODCAST



Baobab tree deaths; zebrafish stem cells; and ice in Antarctica  
[nature.com/nature/podcast](http://nature.com/nature/podcast)

JOHN GIBBENS/ALAMY

## POLICY

# China introduces sweeping reforms against misconduct

Policies include creation of journal blacklist and assigning policing to government agency.

BY DAVID CYRANOSKI

China is getting tough on scientific misconduct. The country's most powerful bodies, the Chinese Communist Party and the State Council, introduced a raft of reforms on 30 May aimed at improving integrity across the research spectrum, from funding and job applications to peer-review and publications.

Under the new policy, the Ministry of Science and Technology (MOST) will be responsible for managing investigations and ruling on cases of scientific misconduct, a role previously performed by individual institutions. And for the first time, misconduct cases will be logged in a national database that is currently being designed by MOST.

Inclusion in the list could disqualify researchers from future funding or research positions, and might affect their ability to get jobs outside academia. The Chinese Academy of Social Sciences will oversee the same process for social scientists.

The policy also states that MOST will establish a blacklist of 'poor quality' scientific journals, including domestic and international titles. Scientists who publish in these journals will receive a warning, and those papers will not be considered in assessments for promotions, jobs or grants. A couple of such blacklists already exist, but rarely are they run formally by a government agency.

"Making it clear that articles published in 'bad' journals won't count towards assessment of performance sends a strong signal," says Paul Taylor, who heads a scientific-integrity programme at RMIT University in Melbourne. The plan to crack down on poor-quality and predatory journals is a good idea in practice, he says, but the ministry could find it difficult to identify problematic journals because some are more obvious than others — a challenge that curators of other blacklists have experienced. "It will be interesting to see the criteria that are developed to make these assessments," he says.

A start date for the reforms is yet to be announced, but is expected soon. Researchers in China and abroad say the policy will have considerable impact. "These new rules will make a major difference over time," says Xue Lan, a science- and innovation-policy researcher at Tsinghua University in Beijing.

Scientific misconduct is a significant



The Chinese Academy of Sciences will work with the science ministry to set standards.

problem in China, which has seen a steady stream of plagiarism cases, uses of fraudulent data, falsified CVs and fake peer reviews.

## HELD TO ACCOUNT

Xue says the reforms are more practical than previous policies, which were based on general principles, such as improving researcher ethics, and were therefore hard to implement. "They lay out an accountability system in a detailed way that has never been seen before," he says.

As part of the reforms, the science ministry will work with agencies such as the Chinese Academy of Sciences to create standards for determining misconduct, protocols for monitoring and investigating allegations, and rules for deciding on the severity of penalties according to the type of misconduct. The policy states that funding and jobs can be revoked. Although universities currently have these powers, some scientists say they are rarely applied. "The life-long accountability system will make everyone afraid to commit academic misconduct," says Yu Hailiang, a mechanical engineer at Central South University in Changsha, who blogs about science-integrity issues. The rules will help to establish a good academic atmosphere, he says.

Science-policy researcher Tang Li from Fudan University in Shanghai also supports the reforms, although she worries that, if penalties are too harsh, it might prompt a backlash from researchers. She also warns

that the ministry will need to protect whistleblowers and researchers who are wrongly accused — something that is happening more now that researchers can publicize accusations against each other online. Taylor would like to see some assurance that the investigative process that led to people being named on a misconduct database was fair and rigorous.

The new rules also state that institutions could have their funding revoked if they protect researchers who have conducted serious misconduct. Nicholas Steneck, a researcher in scientific integrity at the University of Michigan in Ann Arbor, says the plan to punish institutions and journals is unique, and the journal policy could be a model for other countries to follow.

The policy also includes a plan to overhaul how researchers are evaluated for jobs and research grants. The current system places significant weight on the number of papers a scientist has published. But some researchers have noted that this encourages corner-cutting and fraud. The new rules call instead for universities instead to consider quality as well as quantity, and to focus on overall innovation and impact as well as publication record.

Efforts to change the culture of science in China will be key to reducing misconduct, says Yu. "As time goes on, these rules will go deep into the hearts of every researcher, allowing people to consciously resist academic bad habits." ■





Virologists testing samples for Ebola in Liberia in 2014.

Most Ebola lab diagnostics detect genetic sequences that are specific to the virus in blood, serum and other bodily fluids. During the West African epidemic, these manual tests required highly trained scientists working in sophisticated tight-security labs that were often far from outbreak zones.

In the DRC, by contrast, the same tests are automated and are being performed more rapidly and nearer to transmission zones by GeneXpert. The machine uses custom cartridges for different diseases and was developed for resource-poor settings. In response to the 2014–16 epidemic, the machine's manufacturer developed a cartridge called Xpert Ebola to test for the Zaire strain of Ebola, which is behind the current DRC outbreak. The turnaround time from taking a sample to receiving a diagnosis in this outbreak is usually a matter of hours or at most a day, Perkins says.

The DRC outbreak, which is mainly centred in remote regions of Équateur Province in the northwest of the country, is still relatively small, with 38 lab-confirmed cases, 14 probable cases and 14 suspected (see 'Rapid response'). As of 9 June, 28 of the people thought to have the virus had died.

Health officials are cautiously optimistic that the outbreak can be stopped quickly. But the WHO is still worried that the virus could spread across the DRC and the rest of Central Africa. Some cases may have gone undetected, which could lead to resurgences of the disease. And for the first time in the DRC, cases have occurred in an urban area.

Should further regional outbreaks occur, GeneXpert machines will continue to be a big help, officials say. There are already around 150 of the machines in the DRC and several hundred in nearby countries for testing for tuberculosis and other diseases. By swapping in Xpert Ebola cartridges, a large Ebola-testing network could be quickly created, says Perkins.

Thanks to lessons learnt from the West African outbreak, says Mara Jana Broadhurst, an Ebola-diagnostics specialist at Stanford University in California, "a new paradigm for Ebola virus detection and diagnosis is taking shape". ■

## PUBLIC HEALTH

# Fast Ebola test limits outbreak

*Health workers in the Democratic Republic of the Congo can diagnose the virus in hours, instead of days.*

BY DECLAN BUTLER

Health workers fighting the Ebola epidemic that swept West Africa several years ago waited days, even a week, for the results of laboratory tests to detect the deadly virus. But in an Ebola outbreak that began in early April in the Democratic Republic of the Congo (DRC), this waiting time has shrunk to hours — thanks to a genetic test that was developed in response to the 2014–16 West African epidemic.

Researchers and health officials credit the faster tests with helping to contain the spread of Ebola in the DRC, by allowing infected people to be isolated and their contacts traced promptly. And should sparks from this outbreak light new fires in neighbouring countries, the nimbler test could help to avert a repeat of the devastating West African epidemic.

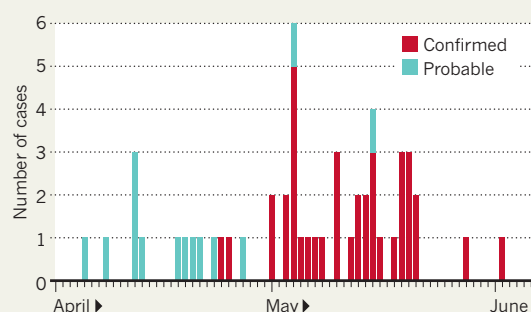
The test involves a small machine called the GeneXpert, which is widely used across Africa to diagnose tuberculosis. The DRC's government has made the GeneXpert its primary method of testing for Ebola in the current outbreak. "Labs have been set up with two to three days' notice in new transmission zones, whereas in West Africa it took months and months to get facilities up and running," says Mark Perkins, head of laboratory networks

for the World Health Organization (WHO) Health Emergencies Programme. "It's a remarkable change."

One of the biggest lessons of the West African epidemic — in which around 29,000 people were infected and 11,000 died in Sierra Leone, Guinea and Liberia — was the need to diagnose cases of Ebola more quickly. Better lab tests could have averted 30–70% of cases and saved thousands of lives and billions of dollars, according to a May 2018 report by the Foundation for Innovative New Diagnostics in Geneva, Switzerland.

## RAPID RESPONSE

With the GeneXpert test, health workers fighting the current outbreak of Ebola in the Democratic Republic of the Congo have rapidly set up testing labs in transmission zones to deliver fast results.



SOURCE: WHO

## GEOENGINEERING

# Price of sucking CO<sub>2</sub> from air plunges

*Technology moves closer to economic viability.*

BY JEFF TOLLEFSON

**S**iphoning carbon dioxide from the atmosphere could be more than an expensive last-ditch strategy for averting climate catastrophe. A detailed economic analysis published last week suggests that the geoengineering technology is inching closer to commercial viability.

The study was conducted by researchers at Carbon Engineering in Calgary, Canada, which has been operating a pilot CO<sub>2</sub>-extraction plant in British Columbia since 2015. That plant — based on a concept called direct air capture — provided the basis for the economic analysis, which includes cost estimates from commercial vendors of all of the major components (D. W. Keith *et al.* *Joule* <http://doi.org/cqj>; 2018).

Depending on a variety of design options and economic assumptions, the cost of pulling 1 tonne of CO<sub>2</sub> from the atmosphere ranges from US\$94 to \$232. By contrast, the previous comprehensive analysis of the technology, conducted by the American Physical Society in 2011, estimated that it would cost \$600 per tonne (see [go.nature.com/2xuauq7](http://go.nature.com/2xuauq7)).

Carbon Engineering, which was founded in 2009, says that it published the paper to advance discussions about the approach's cost and potential. "We're really trying to commercialize direct air capture in a serious way," says David Keith, the company's acting chief scientist and a climate physicist at Harvard University in Cambridge, Massachusetts.

"It's great to see human ingenuity marshalling around a problem that at first pass seemed

to be intractable," says Stephen Pacala, co-director of the carbon-mitigation initiative at Princeton University in New Jersey. He gives the Carbon Engineering team credit for publishing its results and subjecting its proprietary technology to public scrutiny.

The company's design blows air through towers that contain a solution of potassium hydroxide, which reacts with CO<sub>2</sub> to form potassium carbonate. The result, after further processing, is a calcium carbonate pellet that can be heated to release the CO<sub>2</sub>. That CO<sub>2</sub> could then be pressurized, put into a pipeline and disposed of underground, but the company is planning instead to use it to make synthetic, low-carbon fuels. Keith says that Carbon Engineering can produce these for a cost of about \$1 per litre. When the company configured the air-capture plant for this purpose, it was able to bring costs down to as low as \$94 per tonne of CO<sub>2</sub>.

Klaus Lackner, a pioneer in the field who heads Arizona State University's Center for Negative Emissions in Tempe, says that Carbon Engineering has taken a "brute-force" approach to driving down costs using known technologies. "They are coming within striking distance of making this interesting economically," he says. ■

## MICROBIOLOGY

# Faecal transplants could help preserve vulnerable species

*New gut bacteria can expand the diets of animals such as koalas and rhinoceroses.*

BY SARA REARDON

**K**oalas are among the world's fussiest eaters, consuming only the leaves of eucalyptus trees — and just a few varieties of eucalyptus at that. Research now suggests that the animals' discriminating diet is determined in part by the bacteria that live in their guts, which seem to restrict an individual koala's ability to digest certain species of eucalyptus.

The finding, which was presented on 8 June at the annual meeting of the American Society for Microbiology (ASM) in Atlanta, Georgia, comes amid a growing interest in how an animal's microbiome influences its ability to adapt to environmental change. Scientists studying koalas and other vulnerable species are trying to find out whether altering an animal's gut bacteria through its diet — or even faecal transplants — can increase survival.

That is an urgent question for the koala (*Phascolarctos cinereus*), whose habitat in

Australia is under threat from human activity. In some places, the koala population dwarfs the supply of eucalyptus — but even when the animals are transplanted to areas with abundant food, some die. Experiments by koala ecologist Ben Moore and his colleagues at Western Sydney University in Australia suggest that this might be due to an incompatibility between available eucalyptus varieties and the mix of an individual koala's gut bacteria.

Moore and his colleagues collected faeces from 200 koalas at 20 sites around Australia. When the researchers analysed the plant materials in the faeces, they found that some koalas ate only a highly nutritious eucalyptus species known as manna gum (*Eucalyptus viminalis*). Others ate less-nutritious messmate (*E. obliqua*), and only a fraction of the animals would eat both — even at the same site.

When Moore and his colleagues analysed the microbial make-up of the faeces, they found that the koalas that preferred manna-gum eucalyptus harboured different bacteria

from those that ate messmate. In an attempt to test whether the different diets were the cause or the result of the different microbiomes, the researchers transplanted faeces from six wild koalas that ate messmate into six wild koalas that preferred manna gum. Within 18 days, the microbiomes of the koalas that underwent the procedure were nearly identical to those of the donor animals. A few of the animals that received transplants also seemed more willing to eat messmate.

To Moore, this suggests that koala-to-koala faecal transplants might help to expand the types of food available to individual animals, and increase their chances of survival. Eria Rebollar, a microbial ecologist at the National Autonomous University of Mexico in Mexico City, says that the koala study is one of the first demonstrations that faecal transplants can modify wild animals' microbiomes.

Other experiments suggest that some animals could benefit from having their microbiomes reshaped by faecal transplants ►





Koalas can be very picky about what kind of eucalyptus they eat.

► from another species. A team led by Denise Dearing, a molecular biologist at the University of Utah in Salt Lake City, found last year that desert woodrats (*Neotoma lepida*) — distant relatives of laboratory rats — carry gut bacteria that allow them to eat plants containing oxalate, a chemical that causes kidney stones. When the scientists, who work with Moore's

rats, the lab rats gained the ability to degrade oxalate.

In some cases, helping endangered species survive might require changing their diets to accommodate their existing microbiomes. At the ASM meeting, scientists from the San Diego Zoo in California presented findings that suggest how the gut microbiome of the near-threatened southern white rhinoceros

(*Ceratotherium simum simum*) could interfere with its fertility. Captive-born southern white rhinos do not reproduce well. GETTY

The team from the San Diego Zoo compared the faeces of captive white rhinos with those of one-horned rhinos (*Rhinoceros unicornis*), which reproduce well in captivity. The white-rhino faeces contained chemicals known as phytoestrogens, which are present in some plants and affect female reproductive hormones. Because both species of rhino ate the same diet, the researchers suspected that their gut microbes might break phytoestrogens down differently.

To test this, the zoo workers switched the female white rhinos' diet to grass pellets, which are low in phytoestrogens. Within two years, two females that had never successfully reproduced became pregnant; they later gave birth to healthy calves. Candace Williams, a molecular biologist at the San Diego Zoo, says that the facility is now feeding grass pellets to all its rhinos. She and her colleagues are trying to identify which bacteria might be responsible for the shift.

Dearing predicts that science will soon reveal many more instances of animals' microbiomes affecting their ability to survive. "I think it's more common than we've been able to document," she says. "We just didn't have the tools to do this until recently." ■



# TROUBLE IN THE PERIODIC TABLE

*Scientists are changing the rules for approving new elements in the wake of concerns over four recent discoveries.*

ILLUSTRATION BY KAROL BANACH

**T**he mood at Bäckaskog Castle in southern Sweden should have been upbeat when chemists and physicists gathered there for a symposium in May 2016. The meeting, sponsored by the Nobel Foundation, offered researchers a chance to take stock of global efforts to probe the limits of nuclear science, and to celebrate four new elements that they had added to the periodic table a few months earlier. The names of the elements were due to be announced within days, a huge honour for the researchers and countries responsible for the discoveries.

Although many at the meeting were thrilled with how their field was developing — and the headlines it was generating — a significant number were worried. They feared that there were flaws in the process of assessing claims about new elements, and were concerned that reviews of the recent discoveries had fallen short. Some felt there was not enough evidence to justify enshrining the most controversial elements, numbers 115 and 117. The scientific integrity of the periodic table was at stake.

Towards the end of the meeting, one scientist asked for a show of hands on whether or not they should announce the elements' names as planned. The question exposed the depth of concern among the crowd. Most researchers voted to delay the announcement, says Walter Loveland, a nuclear chemist at Oregon State University in Corvallis. And that triggered a remarkable reaction from some of the Russian scientists who had led efforts that resulted in three of the elements. "They just stomped their

BY EDWIN CARTLIDGE

feet and walked out," says Loveland. "I've never seen that in a scientific meeting."

Despite the concerns, the elements' names were announced soon after. Nihonium (atomic number 113), moscovium (115), tennessine (117) and oganesson (118) joined the 114 previously discovered elements as permanent additions to the periodic table. Nearly 150 years after Dmitri Mendeleev dreamed of this organizational structure, the seventh row of the table was officially complete.

Yet the way in which events played out deeply upset some researchers. Claes Fahlander, a nuclear physicist at Lund University in Sweden, expects that experimental results will eventually support the claims for moscovium and tennessine. Nevertheless, he maintains it was "premature" to approve the elements. "We are scientists," he says. "We don't believe — we want to see proof."

As the world prepares to celebrate the International Year of the Periodic Table in 2019, debate over the four additions has forced reforms to the process for verifying other new elements in the future. And the controversy has cast a cloud of uncertainty over the bottom row of elements: it is possible that the table's governing bodies might reassess some of the latest discoveries.

Part of the controversy stems from a rift between some chemists and physicists over who should be the legitimate custodians of the periodic table. Chemists have historically occupied that role, because they



discovered the naturally occurring elements through chemical techniques over centuries of work.

For the past several decades, however, nuclear physicists have led the hunt for new elements — creating them artificially by smashing atomic nuclei into targets. It can take years to produce just one atom of these superheavy elements, which are also notoriously unstable, splintering through radioactive decay in sometimes fractions of a second. So, as groups have vied to be first to create the next elements, it has become more difficult to establish proof of their discoveries.

## SIBLING RIVALRY

Responsibility for approving or rejecting new elements lies with two sister organizations: the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Pure and Applied Physics (IUPAP). Since 1999, they have relied on the judgement of a panel of experts known as the joint working party (JWP), chaired by Paul Karol, a nuclear chemist and emeritus professor at Carnegie Mellon University in Pittsburgh, Pennsylvania. Re-established periodically to assess claims for discoveries as they arise, the latest version of the JWP assembled in 2012 and disbanded in 2016. It consisted of Karol and four physicists.

During that time, the group awarded credit for the discovery of elements 115, 117 and 118 to a Russian–US collaboration led by veteran nuclear physicist Yuri Oganessian of the Joint Institute for Nuclear Research (JINR) in Dubna, Russia. And the panel assigned element 113 to researchers at the RIKEN Nishina Center for Accelerator-Based Science near Tokyo.

The JWP's decisions were announced publicly on 30 December 2015, when IUPAC issued a press release trumpeting the discoveries of the four new elements (which had not yet received their formal names). Union officials said they had worked quickly to broadcast the decisions. In fact, they made the announcement before the union's executive committee could approve the JWP's conclusions, as is specified by the union's published rules<sup>1</sup>; that approval came the following month. More controversially, the JWP's findings hadn't even been shown to the physics union, IUPAP, which had expected to see them, says Bruce McKellar of the University of Melbourne in Australia, who was IUPAP's president at the time.

That omission inflamed pre-existing tensions between the two unions. Cecilia Jarlskog, a physicist at Lund University and IUPAP president before McKellar, claims that, for years, the chemistry union has unfairly dominated the process of assessing discoveries. (Karol told *Nature* that in preparing the JWP's reports, he liaised almost exclusively with the chemistry union.) Venting her frustration at the 2016 Swedish meeting, she accused IUPAC of trying to steal the limelight by announcing the discovery on its own, and argued that only physicists “have the competence” to assess claims, according to the published version of her presentation<sup>2</sup>.

On this occasion, tensions in the physics and chemistry communities were heightened by criticism over the JWP's assessment of claims for elements 115 and 117. The JWP backed<sup>3</sup> the conclusions of the team that discovered these elements, which found that chains of radioactive decay from elements 115 and 117 matched up in a way that bolsters the evidence for both discoveries. But this kind of ‘cross-bombardment’ analysis is notoriously tricky for odd-numbered elements. Fahlander and his co-workers at the University of Lund report<sup>4</sup> that the match-up is highly unlikely to exist for 115 and 117 — a concern brought to the attention of the JWP in February 2015.

Panel member Robert Barber, a nuclear physicist at the University of Manitoba in Winnipeg, Canada, says that although he and his colleagues “were very concerned” about cross-bombardment, they concluded there was no alternative to this type of evidence, and they reached consensus on all their decisions. Loveland also supports the overall decision. And even if the latest JWP got some details wrong, he says, history shows that

its decisions are unlikely to be reversed.

However, Dubna nuclear physicist Vladimir Utyonkov takes aim at the JWP. Although he disagrees with the Lund group's argument about cross-bombardment and is confident that the Russian–US claim is robust, Utyonkov maintains that the panel lacked “high-level” experts in heavy-element synthesis, and says that its draft reports contained numerous errors. Karol defends the work that he and his colleagues did as part of the JWP, saying that they tried to abide by the published criteria governing the assessment process. Overall, he says, “I believe the committee was extremely comfortable with its report”.

But it seems that most delegates at the 2016 meeting in Sweden were critical of the JWP. David Hinde, a nuclear physicist at the Australian National University in Canberra, asked the 50 or so researchers present whether they thought the panel's findings were “scientifically satisfactory”. He says that he got very few positive replies to that question.

## REVIEW QUESTIONS

Despite the various concerns, IUPAC and IUPAP went ahead in June 2016 and announced the names of the four new elements. McKellar admits that he had doubts about doing so, but says that most of the physicists and chemists he consulted told him that the JWP's overall conclusions — if not all of the details of their analyses — were probably sound.

Jan Reedijk, then president of IUPAC's inorganic-chemistry division, says that the initial announcement was made early to avoid press leaks and to satisfy demands from the claimant labs, which were eager to get the news out. To enable that, he says, he quickly approved the JWP's findings in December 2015 on behalf of his division, after it had been peer-reviewed and accepted for publication in the union's journal *Pure and Applied Chemistry*. “I noted that the proper refereeing had been done, so gave my ‘yes’ in less than an hour,” he says.

However, it is unclear whether a truly independent review took place. According to the chemistry union's executive director Lynn Soby, the JWP's work was reviewed in a two-step process before the announcement. First, its findings went to several labs, principally ones involved in the latest discoveries as well as another reviewer suggested by one of the labs. Then the JWP's reports were sent to members of the chemistry union's committee on terminology, nomenclature and symbols.

Soby says that the committee's job was to check wording and formatting errors, and that therefore it was down to the labs themselves to provide scientific scrutiny. She says that was appropriate, given that they are the experts in that field. Yet one of those researchers, Utyonkov, thought that the chemistry union had recruited 15 independent experts to do the scientific review. He assumed that he and two Dubna colleagues had been asked to check only facts and figures in the reports. “I don't know how we can be considered as independent referees,” he says.

Looking back, Jarlskog wishes that she and the rest of the physics community had paid closer attention to how the entire assessment process was completed, particularly the refereeing of the JWP's conclusions. “I am going to have nightmares about how negligent we have been.”

To address the concerns raised, the two unions have agreed on new procedures for assessing any future elements. According to the amended rules, which were released in May (see [go.nature.com/2ji1gv4](http://go.nature.com/2ji1gv4)), the presidents of IUPAC and IUPAP will now each get the chance to review the JWP's findings before announcing their conclusions together. To do so, they will carry out an independent peer-review process alongside that of *Pure and Applied Chemistry*.

McKellar says that the changes will have a positive effect. “Each union has developed a good bit of trust working together on this,” he says.

But those changes won't satisfy some critics, such as Jarlskog. “I just don't think that the new rules will change anything,” she says. ■

Edwin Cartlidge is a reporter in Rome.

1. Corish, J. *Chem. Intl.* **38**, 9–11 (2016).
2. Jarlskog, C. *EPJ Web Conf.* **131**, 06004 (2016).
3. Karol, P. J., Barber, R. C., Sherrill, B. M., Vardaci, E. & Yamazaki, T. *Pure Appl. Chem.* **88**, 139–153 (2016).
4. Forsberg, U. *et al. Phys. Lett. B* **760**, 293–296 (2016).

# COMMENT

**PANDEMICS** Invest in screening people for infections, not animals for viruses **p.180**



**CONSERVATION** Krill, orca and whales: three tales of ocean plunder **p.184**

**PHYSICS** Has the pursuit of beauty led modern physics into a morass? **p.186**

**OBITUARY** Stanley Falkow, microbe-mechanism hunter, remembered **p.190**

PAUL SUTHERLAND/NATIONAL GEOGRAPHIC/GETTY



Wandering albatrosses follow a vessel as it fishes for toothfish.

## Watch over Antarctic waters

In a rapidly changing climate, fisheries in the Southern Ocean must be managed cautiously in response to data, warn **Cassandra Brooks** and colleagues.

Antarctica is a “natural reserve, devoted to peace and science”, according to the Antarctic Treaty System. This complex set of agreements collectively takes a firm stance on conservation, exemplified by the Convention on the Conservation of Marine Living Resources. Adopted in 1980, this convention was negotiated rapidly in response to expanding trawling of Antarctic krill (*Euphausia superba*). Krill are at the base of the region’s marine food web, so there were worries that a dearth of the small crustaceans would threaten the whole ecosystem, especially whales.

The aim of the convention is to conserve all biota and ecosystems in the Southern Ocean. Although fishing is allowed, it is not a right and does not trump responsibility for conservation. The convention’s provisions

are strict, precautionary and science-based. Nations that are signatories must avoid significant or irreversible damage to fish and other animals that depend on them.

But the convention is failing to protect the Southern Ocean from overfishing and the impacts of climate change.

Up to 20 nations fish in these icy waters<sup>1</sup>. Antarctic krill and Patagonian and Antarctic toothfish (*Dissostichus eleginoides* and *Dissostichus mawsoni*) are the main quarry. More vessels and more-efficient fishing technologies are now able to catch more animals (see ‘Antarctic fisheries’). Vessels using vacuum pumps can suck up 800 tonnes of krill in one day<sup>2</sup>. The vessels compete with birds and mammals for food, especially in the most accessible waters.

At the same time, ocean temperatures,

currents and weather patterns are changing<sup>3</sup>. The northwest coast of the Antarctic Peninsula is one of the fastest-warming places on Earth — summer mean temperatures are on average 3 °C higher than they were in 1950. Diminishing sea ice also means fewer algae, krill and Antarctic silverfish (*Pleurogramma antarctica*). Cumulative impacts of historical and current fishing combined with environmental change have been linked to declines in populations of Chinstrap<sup>4</sup> and Gentoo penguins<sup>5</sup> (*Pygoscelis antarctica* and *Pygoscelis papua*).

Because of the convention’s strict provisions, its 25-member implementing body — the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) — is widely seen as a leader in high-seas fisheries management. ►



► But some fishing states are now trying to weaken the convention's rules. China, CCAMLR's newest member (as of 2007), argues that the convention enshrines nations' rights to fish, rather than a responsibility to conserve<sup>6</sup>. China has also insisted that no-fishing zones are contrary to the convention, even though they are expressly included. And it has proposed that scientific evidence of a threat is required before an area is closed to fishing. Some other fishing nations, such as Russia, support this view<sup>7</sup>.

Because CCAMLR operates by consensus, any state can block a measure that it perceives is not in its interests. For instance, in 2011, South Korea prevented the black-listing of one of its vessels that was caught fishing illegally<sup>8</sup>.

If fishing carries on at its current pace amid rapid climate change, prime Antarctic fisheries and marine ecosystems could collapse, as has happened elsewhere<sup>9</sup>. For example, in the 1990s, after politicians failed to act on scientists' warnings, the abundance of Atlantic cod (*Gadus morhua*) declined to less than 1% of historical levels<sup>9</sup>.

We urge CCAMLR to better protect fisheries in the Southern Ocean. The impacts of climate change on populations now and in the future should be factored into decision-making, to avoid crashes in populations. CCAMLR may need to reduce or stop fishing in threatened areas or where there is high uncertainty about adverse effects. Marine reserves must be well designed, and more of them must be implemented.

CCAMLR should also do more to support basic research that is independent of the fishing enterprise. Such studies will lead to greater understanding of the dynamics of targeted species and their vulnerabilities to environmental change and overfishing.

## FISHING PRESSURE

Antarctic waters have long been plundered (see also page 184). Species driven almost to extinction include elephant seals (*Mirounga leonina*), blue whales (*Balaenoptera musculus*), king penguins (*Aptenodytes patagonica*) and marbled rockcod (*Notothenia rossii*)<sup>10</sup>. Some have bounced back; others haven't, such as the rockcod. Even so, remoteness and harsh conditions have protected animals in the seas around Antarctica, in comparison with those elsewhere.

In the Southern Ocean, more krill are caught than any other species (by weight). About 300,000 tonnes are caught annually. They are mainly destined for omega-fatty-acid supplements and fishmeal. Most krill are caught off the Antarctic Peninsula. The industry says that such catches are small, compared with the more than 300 million tonnes of krill estimated to reside in circumpolar waters<sup>2</sup>.

Patagonian and Antarctic toothfish each support relatively small fisheries in the

An emperor penguin and an ice-breaker in the Ross Sea.



Southern Ocean (see 'Antarctic fisheries'). Owing to high prices, this is lucrative. Exploitation rocketed in the 1990s, when toothfish, rebranded as Chilean sea bass, became popular in top restaurants. Illegal, unreported and unregulated fishing soared and ravaged populations; illegal fishers took six times more fish than did legal vessels<sup>11</sup>. CCAMLR turned this situation around by documenting catches, monitoring vessels and black-listing those that did not comply. Illegal catches fell from 33,000 tonnes in 1996 to less than 2,000 tonnes by 2007 (ref. 11). Nonetheless, many Patagonian toothfish populations crashed, and remain depleted, notably those around the Prince Edward Islands, BAZZARE Bank and Kerguelen Plateau.

The life cycles of toothfish make them particularly vulnerable, as well as difficult to study. They mature late, grow slowly and can live for 50 years. No one knows how many there are in the Ross Sea, the main international fishery for toothfish. Nor does anyone know when, where or how often they spawn<sup>9</sup>. They are the top fish predator in the Southern Ocean, and they are also key prey of Weddell seals (*Lep-*tonychotes weddellii**) and killer whales (*Orcinus orca*), and compete for smaller fish with Adélie penguins (*Pygoscelis adeliae*).

Fishing states want more. Russia is trying

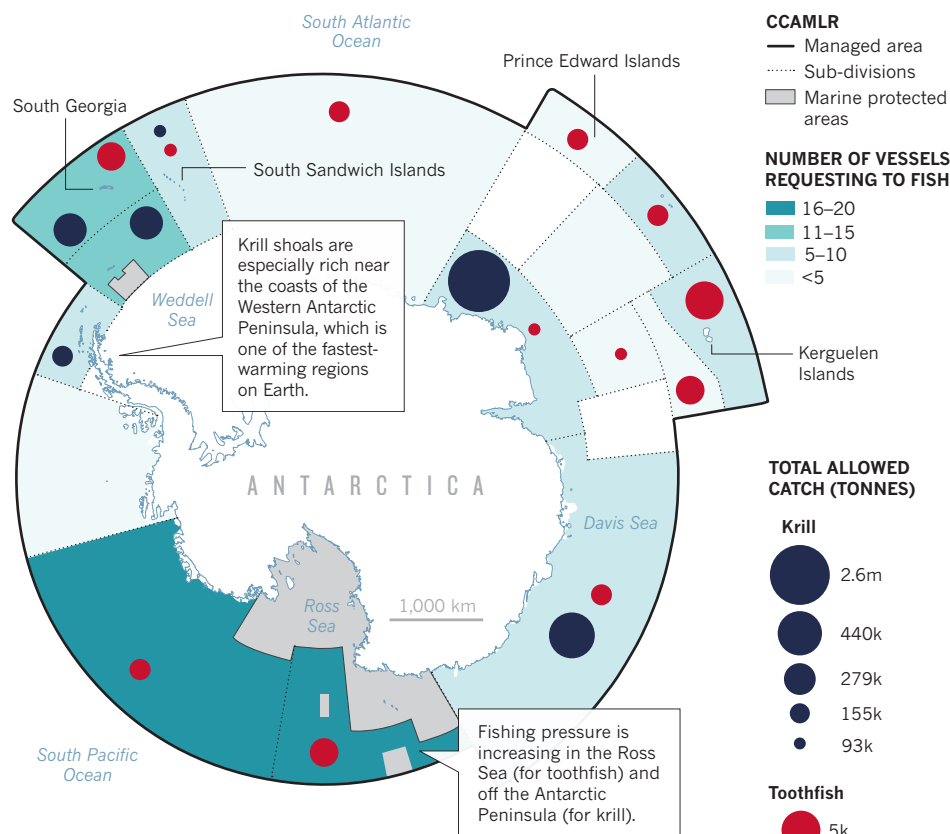
to increase its toothfish catch and send its fleets into unfished areas such as the Weddell Sea<sup>5</sup>. Ukraine wants to capture more krill<sup>5</sup>. New Zealand and Australia, among others, have extended their reach into toothfish areas<sup>1</sup>. Other nations, including Namibia and Uruguay, signed the convention to gain fishing and market access<sup>1</sup>.

## KNOWLEDGE GAPS

Climate change compounds the problem. But environmental effects are difficult to disentangle from the consequences of fishing. For example, scientists do not know whether toothfish fishing or encroaching ice is behind the changing prevalence of killer whales in the southern Ross Sea<sup>12</sup>.

More data would help. But research in the Southern Ocean is difficult and expensive. Much of what is known about toothfish is gathered by the fishing industry, which does not collect environmental data. There are many gaps. We have much to learn about the life histories and population dynamics of the species being caught and how environmental changes affect their birth and death rates. There are few quantitative studies of connections among targeted organisms in the food web.

The long-term ecological research (LTER) programme, based at Palmer Station on the west of the Antarctic Peninsula, is unparalleled in its multi-faceted approach. It gathers information, for example, on how fluctuating sea ice influences krill and other small organisms. CCAMLR also has an ecosystem



SOURCE: WWW.CCAML.R.ORG

monitoring programme designed to study krill-fishing impacts on land-breeding marine predators. However, these data are not effectively incorporated into decision-making.

CCAMLR acknowledges that it must account for the impacts of global warming in its policies. In 2017, it produced a Climate Change Response Work Programme to specify research and monitoring requirements and potential actions<sup>13</sup>. It says it will engage climate-change experts. But progress is slow, and climate change is not. CCAMLR's rules and catch allocations are still based on models that do not consider climate-change scenarios.

### ADAPTIVE ANALOGUES

Responsive, ecosystem-based fisheries management is still being developed. Lessons are emerging from around the world. CCAMLR already has the policy tools to benefit.

Adaptive management is in place off the US West Coast. Since 2015, on the basis of stock surveys and climate indicators, the Pacific Fisheries Management Council has temporarily banned fishing of sardines (*Sardinops sagax*). The closure of this large fishery (109,000 tonnes in 2012) is bringing hardship now. But higher levels of fishing may be possible in future when indicators allow.

A moratorium on fishing in the polar waters of the Arctic is relevant to the Southern Ocean. The 2.8-million-square-kilometre area was not fished before 2017 because it was frozen for much of the year. Now, as around the Antarctic Peninsula, reductions in summer ice

## ANTARCTIC FISHERIES

Krill and toothfish received the largest catch allowances in the Southern Ocean in 2017–18. They are increasingly exploited in spite of tight management by the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR).

are making the area more accessible. In 2017, states bordering the Central Arctic Ocean adopted the Arctic agreement, pledging not to fish there for 16 years, to allow scientists to study of the impacts of climate change first.

In other words, countries can come together to protect sensitive fisheries and the environment, and to support research. The moratorium came about because bordering nations were concerned about over-harvesting. For example, Bering Sea pollock (*Gadus chalcogrammus*) has yet to recover from stock depletion in the 1980s and 1990s. This prompted 2,000 scientists to petition for a fishing ban in 2012. Three years later, the ban was implemented by the United States, Norway, Denmark, Canada and Russia. China, Japan, Korea, Iceland and the European Union have joined since.

Such multi-national cooperation bodes well for the development of strategies for future Arctic fishing that are precautionary and ecosystem-based. Most of these countries are also CCAMLR members. Their willingness to show restraint in the face of uncertainty in the north could be paralleled in the south.

### THREE SOLUTIONS

Given the threat posed by climate change, what are the conditions under which

fishing can continue and still meet the precautionary provisions of the convention? To avoid passing tipping points in the marine food web, CCAMLR needs to take the following three steps.

**Implement more and better-designed marine reserves.** CCAMLR has established two, one in the Ross Sea and one in the South Orkney Islands Southern Shelf, off the eastern tip of the Antarctic Peninsula.

Neither includes comparable reference areas for monitoring fishing against environmental impacts. The Ross Sea protections are set to expire 35 years after they began, which is less than the lifespans of many of the animals intended for protection, such as toothfish. Future closures must ban fishing in the most ecologically crucial areas. The protections should last at least as long as the life expectancies of the animals being safeguarded. And they should include comparable reference areas outside the no-fishing zone.

**Incorporate climate-change scenarios into decision rules.** Current management measures, including catch quotas, are based on models that do not include climate-change scenarios. An environmental shift could cause a population crash in the harvested species or some other species in its food web. To protect against these crashes — and to comply with the provisions



of the convention — CCAMLR must be more precautionary and adaptive. This may mean that quotas are reduced, or that allocations are more temporally and spatially explicit. If the threat of overfishing is deemed readily apparent, or if the level of uncertainty is too high, then CCAMLR may need to temporarily close regions of the Southern Ocean to fishing.

**Develop more-robust research and monitoring programmes.** The Scientific Committee on Antarctic Research (SCAR) should first compile the available information and ongoing research regarding the effects of climate change and fish populations in Southern Ocean ecosystems. The committee undertook these analyses for krill, before establishing the CCAMLR convention. SCAR should then work with CCAMLR scientists, independent experts and non-governmental organizations to identify crucial questions, and what is required to answer them. CCAMLR needs to be more transparent and to invite SCAR and other independent experts into its scientific working groups, from which they are currently excluded.

Governments that are part of CCAMLR will need to fund the research and monitoring efforts, which must be independent of the fishing industry. The Palmer LTER programme shows that the techniques are available, but investment is needed to expand the scientific reach.

CCAMLR states have acted quickly in the past, but change is accelerating in the Southern Ocean. Countries must rise swiftly to this challenge. ■ [SEE INSIGHT P.199](#)

**Cassandra M. Brooks** is an assistant professor in the Environmental Studies Program, University of Colorado Boulder. **David G. Ainley, Peter A. Abrams, Paul K. Dayton, Robert J. Hofman, Jennifer Jacquet, Donald B. Siniff.**

e-mail: [cassandra.brooks@colorado.edu](mailto:cassandra.brooks@colorado.edu)

- Brooks, C. M. *Polar J.* **3**, 277–300 (2013).
- Nicol, S., Foster, J. & Kawaguchi, S. *Fish Fish.* **13**, 30–40 (2012).
- Jacobs, S. *Phil. Trans. R. Soc. A* **364**, 1657–1681 (2006).
- Trivelpiece, W. Z. et al. *Proc. Natl Acad. Sci. USA* **108**, 7625–7628 (2011).
- CCAMLR. *Report of the Thirty-Fifth Meeting of the Commission* (CCAMLR, 2016).
- Jacquet, J., Blood-Patterson, E., Brooks, C. & Ainley, D. *Marine Pol.* **63**, 28–34 (2016).
- Liu, N. & Brooks, C. M. *Mar. Pol.* **94**, 189–195 (2018).
- CCAMLR. *Report of the Thirtieth Meeting of the Commission* (CCAMLR, 2011).
- Abrams, P. A. et al. *Fish Fish.* **17**, 1–23 (2016).
- Ainley, D. G. & Pauly, D. *Polar Rec.* **50**, 92–107 (2013).
- Österblom, H. & Sumaila, U. R. *Glob. Environ. Change* **21**, 972–982 (2011).
- Pitman, R. L., Fearnbach, H. & Durban, J. W. *Polar Biol.* **41**, 781–792 (2018).
- CCAMLR. *Report of the Thirty-Sixth Meeting of the Commission* (CCAMLR, 2017).

# Pandemics: spend on surveillance, not prediction

Trust is undermined when scientists make overblown promises about disease prevention, warn **Edward C. Holmes, Andrew Rambaut and Kristian G. Andersen.**

**T**he resurgence of Ebola virus in the Democratic Republic of the Congo this May is a stark reminder that no amount of DNA sequencing can tell us when or where the next virus outbreak will appear. More genome sequence data were obtained for the 2013–16 Ebola epidemic than for any other single disease outbreak. Still, health workers in Mbandaka, the country's northwestern provincial capital, are scrambling to contain a growing number of cases.

Over the past 15 years or so, outbreaks caused by viruses such as Ebola, SARS and Zika have cost governments billions of US dollars. Combined with a perception among scientists, health workers and citizens that responses to outbreaks have been inadequate, this has fuelled what

seems like a compelling idea. Namely, that if researchers can identify the next pandemic virus before the first case appears, communities could drastically improve strategies for control, and even stop a virus from taking hold<sup>1,2</sup>. Indeed, since 2009, the US Agency for International Development has spent US\$170 million on evaluating the “feasibility of preemptively mitigating pandemic threats”<sup>1</sup>.

Various experts have flagged up problems with this approach (including the three of us)<sup>3,4</sup>. Nonetheless, an ambitious biodiversity-based approach to outbreak prediction — the Global Virome Project — was announced in February this year, with its proponents soliciting \$1.2 billion in funding from around the world (see ‘High stakes’). They estimate that other mammals and birds contain 1.67 million unknown viruses from the families of viruses that are most likely to jump to humans, and will use the funding to conduct a genomic survey of these unknown viruses, with the aim of predicting which might infect people<sup>1</sup>.

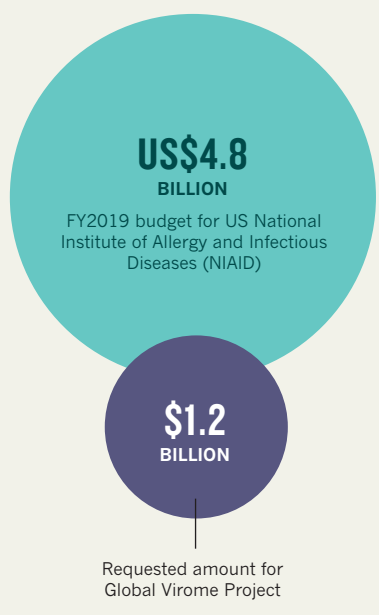
Broad genomic surveys of animal viruses will almost certainly advance our understanding of virus diversity and evolution. In our view, they will be of little practical value when it comes to understanding and mitigating the emergence of disease.

We urge those working on infectious disease to focus funds and efforts on a much simpler and more cost-effective way to mitigate outbreaks — proactive, real-time surveillance of human populations.

The public has increasingly questioned the scientific credibility of researchers working on outbreaks. In the 2013–16 Ebola epidemic, for instance, the international response was repeatedly criticized for being too slow. And during the 2009 H1N1 influenza epidemic, people asked whether the severity of the virus had been overblown, and if the stockpiling of pharmaceuticals was even necessary<sup>5</sup>. Making promises about disease prevention

## HIGH STAKES

Estimated cost of surveying 1.67 million animal viruses is equal to one-quarter of the NIAID's budget for infectious-diseases research.





A health official checks for Ebola symptoms by taking the temperature of passengers arriving at Mbandaka Airport in the Democratic Republic of the Congo.

and control that cannot be kept will only further undermine trust.

### FORECASTING FALLACY

Supporters of outbreak prediction maintain that if biologists genetically characterize all of the viruses circulating in animal populations (especially in groups such as bats and rodents that have previously acted as reservoirs for emerging viruses), they can determine which ones are likely to emerge next, and ultimately prevent them from doing so. With enough data, coupled with artificial intelligence and machine learning, they argue, the process could be similar to predicting the weather<sup>6</sup>.

Reams of data are available to train models to predict the weather. By contrast, it is exceedingly rare for viruses to emerge and cause outbreaks. Around 250 human viruses have been described, and only a small subset of these have caused major epidemics this century.

Advocates of prediction also argue that it will be possible to anticipate how likely a virus is to emerge in people on the basis of its sequence, and by using knowledge of how it interacts with cells (obtained, for instance, by studying the virus in human cell cultures).

This is misguided. Determining which

of more than 1.6 million animal viruses are capable of replicating in humans and transmitting between them would require many decades' worth of laboratory work in cell cultures and animals. Even if researchers managed to link each virus genome sequence to substantial experimental data, all sorts of other factors determine whether a virus jumps species and emerges in a human population, such as the distribution and density of animal hosts. Influenza viruses have circulated in horses since the 1950s and in dogs since the early 2000s, for instance<sup>7</sup>. These viruses have not emerged in human populations, and perhaps never will — for unknown reasons.

In short, there aren't enough data on virus outbreaks for researchers to be able to accurately predict the next outbreak strain. Nor is there a good enough understanding of what drives viruses to jump hosts, making it difficult to construct predictive models.

Biodiversity-based prediction also ignores the fact that viruses are not fixed

***"The challenge is to link genomic, clinical and epidemiological data within days of an outbreak being detected."***

entities. New variants of RNA viruses appear every day. This speedy evolution means that surveys would need to be done continuously to be informative. The cost would dwarf the proposed \$1.2-billion budget for one-time sequencing.

Even if it were possible to identify which viruses are likely to emerge in humans, thousands of candidates could end up being identified, each with a low probability of causing an outbreak. What should be done in that case? Costs would skyrocket if vaccines and therapeutics were proposed for even a handful of these.

### SCREEN AND SEQUENCE

Currently, the most effective and realistic way to fight outbreaks is to monitor human populations in the countries and locations that are most vulnerable to infectious disease. This can be done by local clinicians, health workers in non-governmental organizations such as Médecins Sans Frontières (MSF; also known as Doctors Without Borders), and global institutions such as the World Health Organization (WHO).

We advocate the detailed screening of people who are exhibiting symptoms that cannot easily be diagnosed. Such tests should use the latest sequencing



technologies to characterize all the pathogens that have infected an individual — the human ‘infectome’<sup>8</sup>. To track previous infections, investigators should also assess each person’s immune response, by analysing components of their blood using broad-scale serology<sup>9</sup>.

Emerging diseases are commonly associated with population expansions — when people encroach on habitats occupied by animals — as well as with environmental disturbances and climate change. Deforestation, for instance, can promote human interactions with animals that carry new threats, and can increase encounters with new vector species such as ticks and mosquitoes<sup>10</sup>. Animal die-offs, for example that of bar-headed geese (*Anser indicus*) at Lake Qinghai in China in 2005 (which was caused by the H5N1 influenza virus), can also flag problem regions or emerging pathogens. Surveillance efforts should therefore focus on communities that live and work in such environments.

Identifying which pathogen is causing an outbreak is no longer the bottleneck it once was. It took researchers two years to determine HIV as the cause of AIDS in the early 1980s using microscopy and other techniques. By contrast, in 2012 it took only weeks for investigators using genomic technologies to discover the coronavirus that caused Middle East respiratory syndrome (MERS).

Rapid identification of viruses can be achieved only if such technologies — and the people trained to use them — are globally available, including in resource-limited regions where the risk of outbreaks might be higher. Thankfully, relevant capacity-building programmes are now beginning to be established, such as the Human Heredity and Health in Africa (H3Africa) Initiative, run by the UK Wellcome Trust and the US National Institutes of Health<sup>11</sup>.

Once an emerging outbreak virus has been identified, it needs to be analysed quickly to establish what type it is; which molecular mechanisms (such as receptor type) enable it to jump between individuals; how it spreads through human populations; and how it affects those infected. In other words, at least four kinds of analysis are needed: genomic, virological, epidemiological and clinical. And the data must be passed to key stakeholders, from researchers and health workers on the ground to international agencies such as the WHO and the MSF. Data must be kept as free of restrictions as possible, within the constraints of protections of patient privacy and other ethical issues.

This will best be achieved through an established global network of highly trained local researchers, such as the WHO Global Outbreak Alert and Response Network (GOARN). Real-time tools for



KENNY KATOMBE/REUTERS

People in Mbandaka are taking extra precautionary measures to stop the spread of Ebola virus.

reconstructing and tracking outbreaks at the genomic level, such as portable sequencing devices, are improving fast<sup>8</sup>. Information gathered during recent outbreaks has quickly had tangible impacts on public-health decisions, largely owing to data generation and analysis by many research teams within days of people being infected<sup>12</sup>.

For instance, in the 2013–16 Ebola epidemic, genome sequencing of the virus proved that a person could sexually transmit the disease more than a year after becoming infected. This prompted the WHO to increase its recommended number of tests for persistent infection in survivors of the disease.

Ultimately, the challenge is to link genomic, clinical and epidemiological data within days of an outbreak being detected, including information about how people in an affected community are interacting. Such an open, collaborative approach to tackling the emergence of infectious disease is now possible. This is partly thanks to technology, but is mainly due to a shift in perception about the importance of this approach. At least in genomic epidemiology, there is a growing move towards real-time, open-access data and analysis, aided by the use of preprint servers and wikis such as Virological (<http://virological.org>). This type of collaborative effort can complement the work of agencies including the WHO and the MSF, which focus predominantly on providing information, isolating those who have been infected, and so on.

**“There is a growing move towards real-time, open-access data and analysis.”**

So far, researchers have sampled little of the viral universe. Surveys of animals will undoubtedly result in the discovery of many thousands of new viruses. These data will benefit studies of diversity and evolution, and could tell us whether and why some pathogens might jump species boundaries more frequently than others. But, given the rarity of outbreaks and the complexity of host–pathogen interactions, it is arrogant to imagine that we could use such surveys to predict and mitigate the emergence of disease.

New viruses will continue to emerge unexpectedly. There is a lot we can and must do to be better prepared. ■

**Edward C. Holmes** is professor of biology at the University of Sydney, Australia. **Andrew Rambaut** is professor of molecular evolution at the University of Edinburgh, UK.

**Kristian G. Andersen** is assistant professor of immunology and microbiology at The Scripps Research Institute, La Jolla, California, USA.  
e-mail: [edward.holmes@sydney.edu.au](mailto:edward.holmes@sydney.edu.au)

1. Carroll, D. *et al. Science* **359**, 872–874 (2018).
2. Carroll, D. *et al. Bull. World Health Organ.* **96**, 292–294 (2018).
3. Garrett, L. *Lancet* **391**, 827–828 (2018).
4. Yong, E. ‘Is it possible to predict the next pandemic?’ *The Atlantic* (25 October 2017).
5. Godlee, F. *Br. Med. J.* **340**, c2947 (2010).
6. Morrison, J. ‘Can virus hunters stop the next pandemic before it happens?’ *Smithsonian* (25 January 2018).
7. Parrish, C. R., Murcia, P. R. & Holmes, E. C. *J. Virol.* **89**, 2990–2994 (2015).
8. Gardy, J. L. & Loman, N. J. *Nature Rev. Genet.* **19**, 9–20 (2018).
9. Xu, G. J. *et al. Science* **348**, aaa0698 (2015).
10. Keesing, F. *et al. Nature* **468**, 647–652 (2010).
11. H3Africa Consortium. *Science* **344**, 1346–1348 (2014).
12. Gire, S. K. *et al. Science* **345**, 1369–1372 (2014).





Sperm whales (*Physeter macrocephalus*) socialize in the Indian Ocean.

#### MARINE CONSERVATION

# Sea changes and whale tales

**Sascha Hooker** on three books tackling challenges faced by oceans and marine life.

They cover more than two-thirds of Earth's surface, support much of its biodiversity, produce more than 50% of planetary oxygen and absorb 20–35% of human-created carbon dioxide emissions. Oceans were once considered boundless resources, yet, by 1983, pioneering whale conservationists Stephen Leatherwood and Randall Reeves were writing in *The Sierra Club Handbook of Whales and Dolphins* that “the seas are by no means dead, but they are unquestionably less alive than they were when humanity discovered them”.

Ecological connections between species form a tangled web. Intentional perturbations can lead to unintended consequences, or even shift an ecosystem from one state into another. With the human population soaring towards 8 billion — consuming,

**Spying on Whales: The Past, Present, and Future of Earth's Most Awesome Creatures**  
NICK PYENSON  
Viking (2018)

**The Curious Life of Krill: A Conservation Story from the Bottom of the World**  
STEPHEN NICOL  
Island (2018)

**Orca: How We Came to Know and Love the Ocean's Greatest Predator**  
JASON M. COLBY  
Oxford University Press (2018)

polluting and emitting as it grows — oceans face unprecedented challenges, as do their denizens. Nearly 3 million whales were killed in the twentieth century; the animals must now negotiate hazards such as ship strikes and noise. And global harvests of ocean fish by humans have hit 1 trillion a year.

Now, three books — by palaeobiologist Nick Pyenson, eminent krill scientist Stephen Nicol and historian Jason Colby — explore oceans from the perspectives of whales and krill across palaeontological, decadal and even annual timescales. They examine the evolutionary history and ecology that led to today's ecosystems, and provide insights into ocean-resource management and how changing public sentiment influences the politics behind this.

Pyenson's *Spying on Whales* is a palaeontological howdunnit embedded in a travelogue devoted to chasing living and extinct whales. The author — curator of fossil marine mammals at the Smithsonian Institution in Washington DC — takes us to field sites from Antarctica to Alaska. His narrative captures the excitement of suction-cup tagging of humpback whales,

TONY WU/NATUREPL.COM



and of digs in Panama, seeking answers to deep questions in cetacean science. Why are today's blue whales the largest animals ever to exist? How can past worlds with higher sea levels and more-acidic oceans inform us about the future of today's oceans as climate change takes hold?

Pyenson likens palaeontology to astronomy for its capacity to transport us in imagination to places never experienced. At Cerro Ballena in northern Chile, he uncovers the densest fossil-whale site yet found, with bones dating to between 6 million and 9 million years ago. Once a tidal flat, the site preserved dozens of whale skeletons. Over at least four separate episodes, these whales and marine mammals unique to South America (walrus-faced whales and aquatic sloths), were probably killed en masse by harmful algal blooms.

The ethics of marine-mammal science can be fraught: research merit, benefits to the population and animal suffering must be weighed up. Many in cetacean science consider it unethical to use specimens from commercial whaling operations, arguing that it supports their continuation. Pyenson is one of few researchers today who collect data, as he puts it, at this “uncomfortable intersection”. Yet his passion for research shines through as he describes the discovery of a previously unknown sensory organ between the unfused lower jawbones of rorqual whales, which might help to coordinate the enormous biomechanics of their gulp when feeding. Clearly also a conservationist at heart, he recognizes that whaling has had much broader ecological consequences than originally imagined.

Nicol's *The Curious Life of Krill* takes the reader deep into the Southern Ocean to look at the most abundant multicellular marine organism by weight. Antarctic krill, crustaceans that grow up to 6 centimetres long and feed whales, seals and seabirds, have a 400-million-tonne global biomass. Nicol writes passionately about their biology, exploitation and management. He notes how fisheries have been hampered by krill biology, because the creatures' digestive enzymes cause rapid degradation after death. (Remarkably, these enzymes also have antibiotic properties that could ultimately move krill oil from health food to mainstream medicine.)

Nicol also delves into the influence of whales and whaling on ocean productivity, noting that the twentieth-century decimation of the mammals has not resulted in the krill surplus envisioned. It now seems likely that all whales have a vital role as ecosystem engineers. They transfer nutrients from the depths, where they feed, to surface waters, where their defecation spreads fertilizing iron. That sustains phytoplankton and, in turn, krill. This finding changes how we view marine ecosystems:



*Thysanoessa* krill swarm off California. Krill underpin many marine ecosystems.

no longer as a simple producer–consumer economy, but as complex, interlocking ecological webs. These involve consumption, but also recycling of nutrients and minerals, so that 3D movements of organisms, sediment and suspension are also crucial.

Colby's *Orca* describes a different kind of catch: the killer whale (*Orcinus orca*) live-capture industry in the Pacific Northwest of North America. From 1961 to 1976, 270 of the animals were corralled behind nets and 50 of them were taken for display in aquariums. Colby vividly chronicles a rapid change in public sentiment, from acceptance of shooting killer whales as feared pests in the 1960s to protest and sabotage of capture attempts in the 1970s. Colby, whose father helped with some of the captures, weaves into the narrative first-hand interviews with the saga's main players. The result is immersive and dramatic. Live captures were halted in North America in 1976, and attention then turned to the release of captive whales. However, vast costs and lack of success in freeing the killer whale Keiko (brought to fame in the 1993 film *Free Willy*) suggest that conservation efforts would be better focused on the habitats and prey of remaining wild populations.

Our human tendency to shoot first and ask questions later is evident in whaling, krill fishing and the live capture of killer whales. Consumption is only later followed by research programmes to establish

management regulations. Among those landmark pieces of legislation are the US Marine Mammal Protection Act of 1972 and the 1980 international Convention on the Conservation of Antarctic Marine Living Resources, described by Colby and Nicol, respectively. Both measures are rightly applauded for their ecological, precautionary and data-driven approaches. However, both are still predicated on quotas for removal, or ‘take’. Nicol briefly mentions the focus in recent years on marine protected areas — initiatives that can prohibit fishing and provide refuges for marine species. These present an alternative management tool that should be quantitatively considered in the evaluation and design of what needs protection, and where.

*Spying on Whales*, *The Curious Life of Krill* and *Orca* acknowledge that our oceans are emptier now than they were only 200 years ago. Our ‘shifting ecological baselines’ lead us to value the best of the worst of our impoverished systems today — to marvel at only moderate productivity. Whereas Pyenson and Nicol help to reset our expectations and wake us from our societal and generational amnesia, Colby demonstrates the speed at which societal attitudes can also shift the baseline of our expectations. In this age of extinction, with ongoing changes in ocean chemistry and physics, it is the potential for a sea change in public attitude that presents hope. ■

**Sascha Hooker** is a reader in the Sea Mammal Research Unit of the Scottish Oceans Institute at the University of St Andrews, UK.  
e-mail: s.hooker@st-andrews.ac.uk

**“Our human tendency to shoot first and ask questions later is evident in whaling, krill fishing and the live capture of killer whales.”**





The Large Hadron Collider at CERN has ruled out many elegant physics theories.

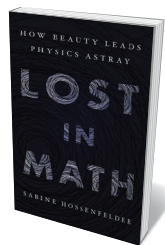
## THEORY

# Beauty, proof and the crisis in physics

Anil Ananthaswamy parses Sabine Hossenfelder's analysis of why the field is at an impasse.

**“W**hy should the laws of nature care about what I find beautiful?” With that statement, theoretical physicist and prolific blogger Sabine Hossenfelder sets out to tell a tale both professional and personal in her new book, *Lost in Math*. It explores the morass in which modern physics finds itself, thanks to the proliferation of theories devised using aesthetic criteria, rather than guidance from experiments. It also charts Hossenfelder's own struggles with this approach.

Hossenfelder — a research fellow specializing in quantum gravity and modifications to the general theory of relativity at the Frankfurt Institute for Advanced Studies in Germany — brings a trenchant new voice to concerns that have been rumbling in physics for at least two decades. In 2006, Lee Smolin's *The Trouble with Physics* and



**Lost in Math:**  
How Beauty Leads  
Physics Astray

SABINE  
HOSSENFELDER  
*Basic* (2018)

Peter Woit's *Not Even Wrong* fired the first salvos at the trend of valuing mathematical elegance over empirical evidence. Both books took on string theory, a 'theory of everything' in which the fundamental constituents of nature are strings vibrating in many more spatial dimensions than the familiar three. Since its entry into mainstream physics in the mid-1980s, the theory has failed to make predictions that would unambiguously verify or falsify it.

Hossenfelder, too, tackles string theory,

but her broadsides are more basic. She points to the paucity of experimental data, exacerbated as the machines needed to probe ever higher energies and smaller distances become more costly to build. Given that, she is worried that too many theorists are using mathematical arguments and subjective aesthetics to judge a theory's validity.

For example, Hossenfelder questions the desire for naturalness — the idea that a theory should not be contrived or have parameters that have to be fine-tuned to fit observations. The standard model of particle physics feels like such a contrivance to many physicists, despite its spectacular success in predicting particles such as the Higgs boson, discovered at the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. In the theory, to prevent the mass of the Higgs from ballooning beyond reasonable bounds, certain parameters have to be set just so, rather than be derived from first principles. This smacks of unnaturalness.

To get rid of this ugliness, physicists developed supersymmetry — an elegant theory in which every known particle has a hypothetical partner particle. Supersymmetry made the Higgs mass natural. It also showed how three of the four fundamental forces of nature would have been one at energies that existed shortly after the Big Bang (an aesthetically pleasing scenario). It even unexpectedly provided a particle, the neutralino, that could explain dark matter — matter that is unseen, yet thought to exist because of its observed gravitational effect on galaxies and galactic clusters. Hossenfelder explains that in combining everything that theoretical physicists value (symmetry, naturalness, unification and unexpected insights), supersymmetry has become “what biologists fittingly call a ‘superstimulus’ — an artificial yet irresistible trigger”.

But despite decades of theorizing by hundreds of top-notch physicists, everyone agrees that supersymmetry is in trouble. The most natural version of it, which requires no fine-tuning, has been ruled out by the LHC data. Hossenfelder quotes theorist Nima Arkani-Hamed as saying that the “best people” were aware of this problem well before the LHC went online. Hossenfelder then chides those very “best people” — no names are given — for not calling “bullshit” on widespread claims that the LHC would discover supersymmetry or dark matter.

Hossenfelder often wears a journalist's hat. Interviews with highly respected physicists (such as theorist Garrett Lisi and Nobel laureates Steven Weinberg and Frank Wilczek) form a significant chunk of *Lost in Math*, as Hossenfelder strives to make sense of the field and her own dissatisfaction with it. We are introduced to myriad problems that plague physics, such as the fine-tuning of the standard model, the lack of a theory of quantum gravity, and



worries about what quantum mechanics is really saying about the nature of reality. (Full disclosure: the latter is the topic of my forthcoming book *Through Two Doors at Once*, which Hossenfelder has endorsed). Hossenfelder also worries about the lack of empirical evidence to sift through in checking the solutions (see N. Wolchover *Nature* 555, 440–441; 2018).

She reproduces many of her discussions with physicists at length, so the material can get a tad repetitive. She could have used her own strong voice to synthesize some of the

**“Despite decades of theorizing, everyone agrees that supersymmetry is in trouble.”**

arguments. Still, there are moments when Hossenfelder’s journalistic forays stand out. For example, her account of meeting the intimidating Weinberg, who

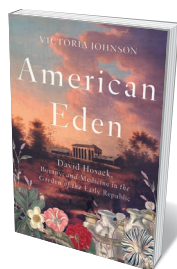
“talks like a book, almost print-ready”, is self-deprecatingly funny and spot-on (I speak from personal experience).

*Lost in Math* is self-aware and dosed with acerbic wit, and it asks bold questions. Hossenfelder’s Twitter followers and readers of her blog, ‘Backreaction’, will recognize her no-holds-barred style. But not all physicists will agree with her. Theorizing in the absence of empirical data is not new, and has paid dividends. For instance, in the early 1960s, the physicist Murray Gell-Mann used symmetry to tidy up the standard model and predict the existence of particles he called quarks. The mathematics turned out to be correct, and he won the 1969 physics Nobel prize for the work. As he noted at the Nobel banquet: “The beauty of the basic laws of natural science, as revealed in the study of particles and of the cosmos, is allied to the liteness of a merganser diving in a pure Swedish lake.”

Hossenfelder acknowledges all this, but she also challenges those who seek to break the current impasse in physics by insisting that nature must be forever beautiful. Admitting that “complaining about aesthetic biases” won’t make the daunting problems in physics go away, she argues for a few ground rules. These include making sure that there is a real problem, which emerges from existing conflicts in theory and data; being clear about one’s assumptions (such as the desire for naturalness or simplicity); and using empirical evidence to choose the right maths for the physics at hand. They are her compass points to prevent us from losing our way in a mathematical jungle, however beautiful. ■

**Anil Ananthaswamy’s next book**, *Through Two Doors at Once*, tells the story of quantum mechanics from the perspective of the double-slit experiment.  
e-mail: anil@nasw.org

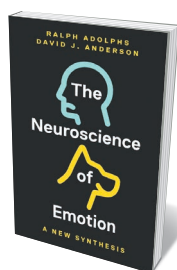
## Books in brief



### American Eden

Victoria Johnson LIVERIGHT (2018)

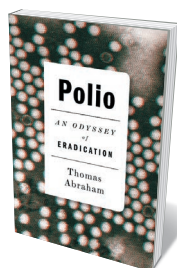
In the 1760s, colonial America was ravaged by yellow fever, typhus and tuberculosis. David Hosack, born into that world, became a titan of medical research in the fledgling nation. He published on tetanus and breast cancer, pioneered smallpox vaccination and, as Victoria Johnson’s fine science biography reveals, contributed vastly to medicinal botany. Hosack’s famed, now lost, Elgin Botanic Garden in New York City became a key training centre for scientists and surgeons, who peered “into the globe-spanning, dizzying complexity of the natural world” through plants. A rich and compelling read.



### The Neuroscience of Emotion

Ralph Adolphs and David J. Anderson PRINCETON UNIVERSITY PRESS (2018)

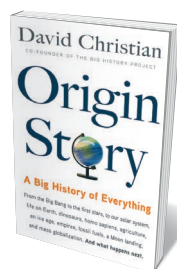
Anger, fear, joy: what are emotions, and what are they for? The sparsity of clear or robust answers spurred neuroscientists Ralph Adolphs and David Anderson to frame an integrated science of emotion. The result is scholarly, lucid and pertinent to both neurobiology and psychology. Mining research from the molecular level to the cognitive, they examine emotions as biological and reflective of evolved adaptations in species as varied as rodents, the fruit fly *Drosophila melanogaster* and *Homo sapiens*. They usefully conclude with open questions for future research.



### Polio: The Odyssey of Eradication

Thomas Abraham HURST (2018)

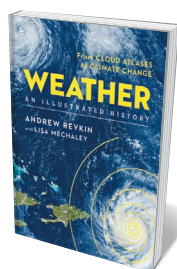
Despite the 99% reduction in polio cases since 1988 under the Global Polio Eradication Initiative (GPEI), the disease lingers on in a handful of countries. Meanwhile, vaccine-derived polioviruses have triggered outbreaks elsewhere. Science journalist Thomas Abraham travelled from slum to boardroom to research the GPEI’s premise and practice, as well as the broader trajectory of the disease and the efforts to tackle it. The result is a trenchant, well-argued analysis, isolating problems such as the initiative’s strategic focus on single vaccinations in regions also riddled with malaria and diarrhoea.



### Origin Story: A Big History of Everything

David Christian LITTLE, BROWN (2018)

Historian David Christian is, with Bill Gates, co-founder of the Big History Project, an online syllabus stretching from the beginnings of the cosmos to human hegemony. Here, Christian distils that 13.8-billion-year chronicle by simplifying the mapping. Each threshold, such as the Big Bang or the lunar landings, is beautifully captured. Heat energy becomes a “drunken traffic cop”; gravity the “virtuoso chain-saw sculptor” of the early Universe; humans uniquely “cultivate and domesticate” information, like farmers. Long-haul science with wit and oomph.



### Weather: An Illustrated History

Andrew Revkin and Lisa Mechaley STERLING (2018)

This fascinating chronicle of humanity’s complex relationship with weather by environmental journalist Andrew Revkin and science educator Lisa Mechaley is told through 100 milestones, each paired with a stunning archival image. A potted history of windscreen wipers sits next to a 1903 schematic of Mary Anderson’s invention; an image of eleventh-century Chinese scientist Shen Kuo faces his prescient observation of climate change; and a dramatic illustration of a waterspout accompanies Benjamin Franklin’s bizarre account of chasing and whipping a whirlwind in 1755. **Barbara Kiser**

# Correspondence

## Wider human-rights focus for health data

Those producing codes of conduct for life-sciences research under the European Union's General Data Protection Regulation (GDPR) should draw on established international work to secure the success of scientific data sharing and the secondary processing of personal data (see also *Nature* 557, 467–468; 2018).

Important guidelines include the 2017 Recommendation on Health Data Governance from the Organisation for Economic Co-operation and Development, and the Framework for Responsible Sharing of Genomic and Health-Related Data developed by the Global Alliance for Genomics and Health (of which we are all members).

The GDPR focuses chiefly on the right to privacy. In addition, the Framework respects the right of everyone “to share in scientific advancement and its benefits” under Article 27 of the Universal Declaration of Human Rights. In our view, this wider human-rights focus will be invaluable in regulating health and genomic research, and the related proportionate interpretation and application of the GDPR.

**Bartha M. Knoppers\*** *McGill University, Montreal, Canada.*  
[bartha.knoppers@mcgill.ca](mailto:bartha.knoppers@mcgill.ca)

*\*On behalf of 5 co-signatories (see [go.nature.com/2lv9d6](http://go.nature.com/2lv9d6) for full list).*

## Don't let the living dead haunt citations

The continued citation of retracted papers — or ‘zombie’ publications — pollutes the scientific literature with fatally flawed studies. The problem is amplified by the common practice of accessing papers through third-party websites such as Google Scholar, ResearchGate and Sci-Hub, which generally do not link to retraction notices. We propose steps publishers could take to prevent new research

from citing retracted studies.

As well as displaying retraction notices more prominently on their websites, journals should post alerts across all pages of the flawed publication. Also, prefacing the paper's title with a notification would warn readers not to download the citation to reference-manager software.

Publishers can ensure that citations of zombie publications are caught before new papers go to press by running automated cross-checks of manuscript reference lists against the Retraction Watch database of retracted papers (<http://retractiondatabase.org>). Universities, too, should ensure that institutional databases are updated to include retraction notices.

**Sandra A. Binning** *University of Montreal, Canada.*

**Fredrik Jutfelt, Josefin Sundin** *Norwegian University of Science and Technology, Trondheim, Norway.*

[sandra.ann.binning@umontreal.ca](mailto:sandra.ann.binning@umontreal.ca)

## Reform Romania's grant-review system

Imagine a Nobel laureate willing to review research-grant applications for Romania. She or he would first need to learn Romanian, to produce a letter of permission to participate from their university president or department chair, and to upload a declaration on Romania's platform for grant reviewers ([www.brainmap.ro](http://www.brainmap.ro)) confirming that they have committed no ethical misdemeanours in the course of their duties in the previous 5 years. These strictures can only lead to the further marginalization and inbreeding of a research system that is hobbled by plagiarism, paltry funds and brain drain.

According to Eurostat, Romania is in the lowest tier of European Union countries in terms of the percentage of gross domestic product (GDP) spent on research and development. In 2018, the nation allocated only

0.18% of GDP to its Ministry of Research and Innovation. About one-quarter of these funds are used in national calls for proposals, which are sparse, unpredictable and currently evaluated by Romanian nationals.

We urge the Romanian government to reinstate the use of international evaluators — with scientific merit as the sole criterion for selection. It should also restore the requirement that proposals be written in English. **Mihai Miclăuș** *National R&D Institute for Biological Sciences, Cluj-Napoca, Romania.*  
**Octavian Micu** *Institute of Space Science, Măgurele, Romania.*  
[mihai.miclaus@icbcluj.ro](mailto:mihai.miclaus@icbcluj.ro)

## Research hotspots in Côte d'Ivoire

Côte d'Ivoire in West Africa has some promising research institutions, despite the considerable social and political turmoil it experienced between 1999 and 2011 (see B. Bonfoh *et al.* *Nature* 474, 569–571; 2011).

The top-ranking institutions for research productivity in 2012–16 in Côte d'Ivoire were its two largest national universities: Félix Houphouët-Boigny University and Nangui Abrogoua University, both in Abidjan (our unpublished results). In third place was the comparatively small Swiss Centre for Scientific Research in Côte d'Ivoire (CSRS), which benefits from a long-standing bilateral research partnership between Côte d'Ivoire and Switzerland. At the CSRS, mutual governance, investment and benefits are balanced with a diverse portfolio of research, education and training within four strategic axes.

In our view, a strong diversification of international funding sources, coupled with an increased share of national government funds, stand to create more centres of scientific excellence in Africa. **Bassirou Bonfoh** *CSRS, Abidjan, Côte d'Ivoire.*  
**Jasmina Saric, Jürg Utzinger**

*Swiss Tropical and Public Health Institute, Basel, Switzerland.*  
[j.saric@swisstph.ch](mailto:j.saric@swisstph.ch)

## Pioneering women in energy physics

I appreciate Roger Fouquet's review of my book *Energy: A Human History*, but take issue with two of his criticisms (*Nature* 557, 162–163; 2018).

First, I did not reference solar-energy pioneer Mária Telkes because her work involved heat storage, not solar electricity — the subject of my discussion.

Second, credit for the discovery of nuclear fission was, in my opinion, more complex than Fouquet implies and not attributable solely to physicist Lise Meitner.

My reading of the history is that the German radiochemists Otto Hahn and Fritz Strassmann discovered fission in Berlin in late 1938. Meitner, who was in Sweden at the time, came up with the physical explanation for this reaction with her nephew Otto Robert Frisch, after corresponding with her former colleagues Hahn and Strassmann. The journal *Naturwissenschaften* (renamed *The Science of Nature* in 2015) received the Hahn–Strassmann paper on 22 December 1938; *Nature* received the Meitner–Frisch paper on 16 January 1939.

It would be odd indeed had I forgotten Meitner after devoting more than 50 pages to her life and work in my 1987 history, *The Making of the Atomic Bomb*. **Richard Rhodes** *Half Moon Bay, California, USA.*

[richardrhodes1@comcast.net](mailto:richardrhodes1@comcast.net)

### CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at <http://go.nature.com/cmchno>.



# Stanley Falkow

## (1934–2018)

Microbe hunter who uncovered how bacteria cause disease.

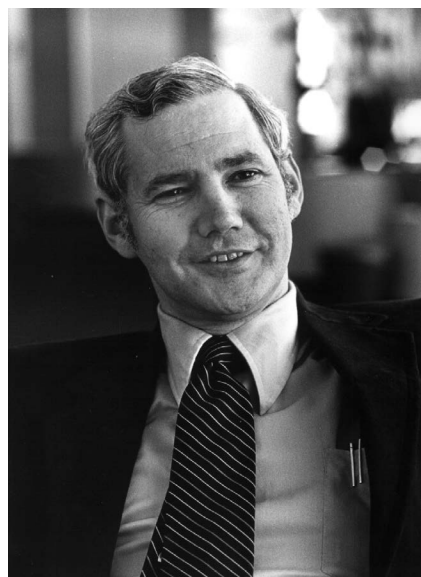
Stanley Falkow discovered the molecular mechanisms through which bacteria cause disease and how to disarm them. But he always described himself as being “on the side of the microbes”.

Falkow, who died on 5 May, began his career in the late 1950s, when DNA was first recognized as the stuff of life, and bacteria were thought to be good model systems for studying it. By the 1960s, he had worked out how to extract and isolate the extrachromosomal packets of bacterial DNA that we now call plasmids, and found that these carry information from one microbe to another. This showed how bacteria can acquire the ability to survive on a new food source or become resistant to an antibiotic.

He made the ominous observation that even harmless bacteria could serve as a source of resistance genes for pathogens. Falkow understood the public-health significance of this, predicted the rise of multidrug-resistant bacteria and campaigned against the overuse of antibiotics. His discoveries of the molecular nature of antibiotic resistance won him the 2008 Lasker–Koshland Special Achievement Award in Medical Science.

The question that framed Falkow's career was: what is a pathogen? Specifically, what makes some microbes disease agents, while others are innocuous or even beneficial? He hypothesized that bacteria have genes coding for specialized ‘virulence factors’ — the transferable, molecular equivalents of claws, fangs, toxins, fur and camouflage — that allow pathogenic bacteria to attack, hide and survive in extreme environments. He spent his life identifying the genes and molecules behind these abilities. He discovered, for example, that some bacteria decorate their surfaces with molecules designed to adhere to host cells, and even fool the host into allowing bacteria to invade. Others have molecular needles that disarm white blood cells by injecting proteins that stop the cells moving, or prevent them from sounding the alarm.

Falkow was born the son of Jewish immigrants in Albany, New York, on 24 January 1934 at the peak of the Great Depression. He was a terrible student, especially in mathematics and science, but loved to read. When he was 11, he found in the local library the book *Microbe Hunters* by Paul de Kruif. Written for lay audiences in 1926, it describes the personalities and adventures of some of the most important figures in microbiology. Falkow devoured it and decided to become a microbe hunter. His life was a conversation



with his heroes from this book.

*Microbe Hunters* begins with Antonie van Leeuwenhoek, who, in the 1670s, was the first to craft glass lenses so powerful that he was able to visualize microscopic organisms. When Falkow read this, he had to have a microscope. He found a Gilbert Company Hall of Science microscope on the shelves of the local toy shop (he could not afford it, but in the end, the shop owner gave it to Falkow). Through this, Falkow witnessed the darting shadows of bacteria growing in the spoilt milk that he kept under his bed.

Falkow's love of microscopy continued throughout his career, which started with a PhD in 1961 from Brown University in Providence, Rhode Island. He learnt electron microscopy and, in the 1970s, at the University of Washington School of Medicine in Seattle, used it to visualize resistance genes that had hopped on to plasmids. In the 1980s at the Rocky Mountain Laboratories in Montana and at Stanford University School of Medicine in California, where he chaired the department of microbiology and immunology, he made micrographs of *Yersinia* and *Salmonella* bacteria invading the intestine.

He was the first to make movies with a videomicroscope to show how *Salmonella* invade gut epithelial cells by causing a literal splash of cytoskeleton and membrane at the cell surface. He adapted diverse methods of painting and tagging molecules inside bacteria and host cells to follow their fate during infection, and even to find bacterial genes that turn on only when inside cells.

Another of Falkow's heroes, Robert Koch, established a set of guidelines in the 1880s (Koch's postulates) to define when a microbe is responsible for an infection. Falkow extended this idea in the 1980s, developing the ‘molecular Koch's postulates’. These describe how one could define virulence at the molecular level — for example by inactivating specific genes to render a microbe harmless, or transferring a gene for adhesion or invasion to an innocuous bacterium to improve its ability to interact with host cells. His lab elucidated the mechanisms of disease of many bacteria (and even yeasts), including major gut pathogens such as *Salmonella* and pathogenic *Escherichia coli*, along with bacteria that cause skin, urinary-tract and sexually transmitted infections, whooping cough, pneumonia, sepsis, and even stomach cancer.

Falkow came to think of the “pathogenic lifestyle” as being not about causing disease, but rather about subtly manipulating a host. He became best friends with his prey. Over time, he spoke more and more about the potential beneficial attributes of bacteria that live on and in our body, foreseeing the study of the human microbiome.

Despite his fame, Falkow often said that his greatest achievements were those of his students. One of his most important gifts was his ability to infect these students with the same sense of awe that he had when he first peered into a microscope. More than 120 scientists, including some of today's most prominent microbiologists, were trained in Falkow's laboratory. Despite his illness, he was still teaching at Stanford last semester, and video-calling into classroom discussions and lab meetings.

For pragmatists interested in tools for conquering the microbial world, Falkow's vision provided the technical framework by which to define potential targets for vaccines and methods to identify, fingerprint and follow the evolution of pathogens through molecular epidemiology. For pure biologists, this same vision provided the foundation for understanding the natural history of pathogenic and mutualistic microbes. For those of us who had the good fortune to know him, he left a bounty of stories and examples of his generosity of spirit as a scientist and a mentor. ■

**Manuel R. Amieva** is an associate professor of Pediatric Infectious Diseases and of Microbiology & Immunology at Stanford University School of Medicine; he was one of Falkow's last postdoctoral fellows.  
e-mail: amieva@stanford.edu

## QUANTUM PHYSICS

# Entanglement on demand

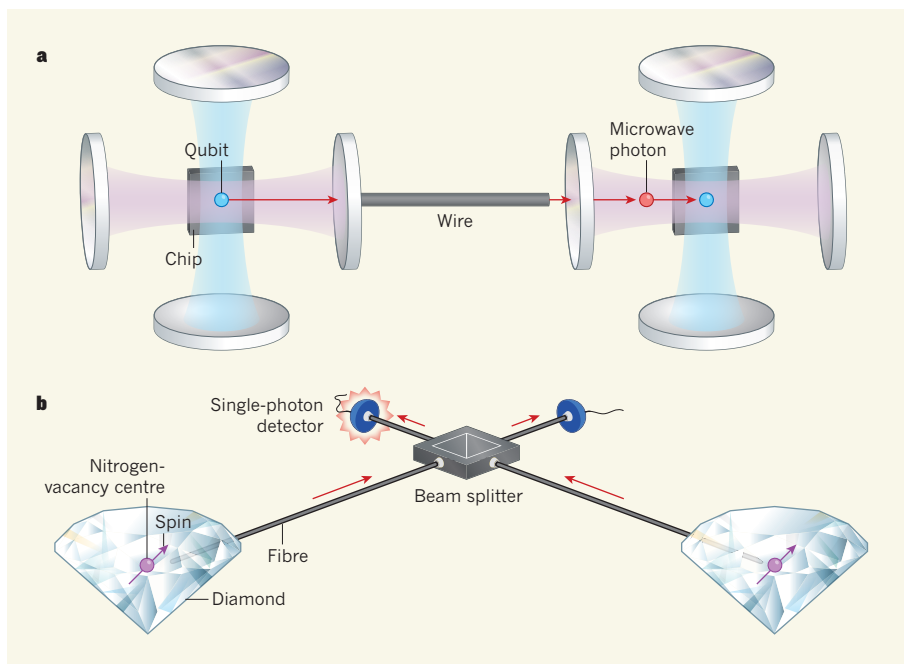
Two experiments show that non-classical correlations can be distributed between distant nodes of a quantum network in a deterministic way. The work smooths the path for extended quantum networks. [SEE LETTERS P.264 & P.268](#)

JULIEN LAURAT

The phenomenon of entanglement, by which two physical systems can be more strongly correlated than is possible in classical physics, is at the heart of quantum mechanics. The efficient distribution of entanglement between distant nodes of a quantum network has become a cornerstone of future quantum technologies — from secure long-distance communication to powerful quantum computing. However, the simultaneous entanglement of more than a few nodes remains a tremendous challenge. On pages 264 and 268, respectively, Kurpiers *et al.*<sup>1</sup> and Humphreys *et al.*<sup>2</sup> demonstrate two methods for delivering entanglement in a deterministic way that could greatly facilitate the construction of large-scale quantum networks<sup>3</sup>.

Quantum nodes take a variety of forms. They can comprise an ensemble of optical emitters, including laser-cooled atoms or ions that are embedded in a crystal. Alternatively, they can rely on a single emitter such as an atom, a defect in diamond or a superconducting quantum bit (qubit). Regardless of the physical platform, the generation of entanglement between two nodes needs to be done in a way that can be extended to many nodes. Probabilistic methods are unsuitable for such a task because simultaneously establishing a large number of entangled nodes leads to an exponential decrease in the overall chance of success.

The challenge of scalability is highlighted, for example, by the prospect of long-distance quantum communication. Unlike the signals in conventional telecommunications, quantum information cannot be amplified or regenerated. As a result, long-distance communication requires a quantum-repeater architecture, in which the distance between the two communicating parties is divided into shorter segments, and entanglement is generated in each segment. The segments can then be connected, enabling long-distance entanglement to be established for use in applications such as quantum teleportation and quantum cryptography. However, if the distribution of entanglement in each segment is probabilistic, the chance of success falls exponentially as the distance increases. This scaling issue is also present in local networks in which many



**Figure 1 | Entanglement delivered on time.** Kurpiers *et al.*<sup>1</sup> and Humphreys *et al.*<sup>2</sup> report two methods for distributing entanglement — or non-classical correlations — between distant nodes of a quantum network in a deterministic fashion. **a**, In Kurpiers and colleagues' experiment, the two nodes consisted of superconducting quantum bits (qubits) that were fabricated on separate chips. Entanglement was generated by a highly efficient process comprising the emission of a microwave photon from one qubit and its capture by the other, through a wire. The authors used devices known as microwave cavities (shown as pairs of disks) to prepare and read out the qubit, and to transfer the photon. **b**, In Humphreys and colleagues' experiment, the two nodes were defects known as nitrogen-vacancy centres in diamond. For each node, the spin (magnetic moment) of the nitrogen-vacancy centre was entangled with the presence or absence of an optical photon emitted into a fibre. The two fibres were connected by a device called a beam splitter, and the detection of a single photon (indicated by the flash) projected the two spins into an entangled state. By producing entanglement more quickly than it was lost, the authors turned a probabilistic protocol into a deterministic one.

modules need to be combined and entangled.

One way to overcome this limitation is to generate entanglement in a deterministic manner: press a button to get entanglement. This is the landmark advance that Kurpiers and colleagues demonstrate. In their experiment, the quantum nodes consisted of two superconducting qubits that were fabricated on separate chips and then connected by a 90-centimetre-long wire (Fig. 1a). Each qubit was strongly coupled to two devices known as microwave cavities. One cavity enabled the preparation and read-out of the qubit, whereas the other facilitated the transfer of microwave photons through the wire.

Kurpiers and colleagues' demonstration

relied on a 20-year-old seminal proposal for directly transferring a quantum state from a sender to a receiver<sup>4</sup>. The authors began by preparing one qubit in a specific state. They then applied microwave pulses to the qubit, which caused it to release a photon that was captured by the other qubit. Because the processes of photon emission and capture can be extremely efficient in superconducting chips, this set-up leads to the distribution of entanglement every 0.02 milliseconds, with an overall fidelity (a parameter that measures how close the entanglement is to the ideal state) of close to 80%. Improvements to the experiment could push this to more than 90%, and error-correction procedures might improve



this number even further. Such on-demand generation of entanglement was also reported recently<sup>5</sup> using a different chip implementation from that of Kurpiers and colleagues.

The direct-transfer strategy has also been carried out using optical photons between distant atoms or ions<sup>6</sup>. In that case, reaching high efficiencies of photon emission and capture remained a challenge. But another way to overcome the prohibitive scaling of quantum networks is to turn a probabilistic method into a deterministic one.

Under a probabilistic protocol, each attempt to produce entanglement has a low chance of success. However, if enough attempts are realized in a given length of time, the generation of entanglement can be ensured in this time frame. Such an approach can therefore provide deterministic entanglement at a predetermined time. But there is a crucial requirement for achieving this goal: entanglement must be produced more quickly than it is lost, or else the generated entanglement could be gone before it has been delivered.

In 2015, a trapped-ion experiment succeeded in breaking this threshold<sup>7</sup>. Humphreys and colleagues have now achieved the same feat using a solid-state system. In their work, the two quantum nodes were single defects, known as nitrogen-vacancy centres, in diamond (Fig. 1b). The authors placed diamonds in cooling devices that were separated by a distance of 2 metres. They then generated entanglement between the spins (magnetic moments) of the nitrogen-vacancy centres by adopting a 'heralded' technique that has been used for other platforms and has enabled rudimentary versions of quantum-repeater segments<sup>8,9</sup>.

Humphreys *et al.* prepared the two nodes so that they had an identical spin state. For each system, the authors used laser pulses to generate entanglement between the spin of the nitrogen-vacancy centre and the presence or absence of an optical photon emitted into a fibre. The fibres from both systems were connected, and the detection of a single photon midway between the nodes projected the two spins into an entangled state. This is because there was no way of knowing, even in principle, from which node the photon was emitted. This single-photon scheme was one of two key ingredients of the authors' work: it enabled a much higher rate of entanglement production than could be achieved in a previous study<sup>10</sup> using nitrogen-vacancy centres that relied on a two-photon process.

The second key ingredient was a dramatic extension in the lifetime of the stored entanglement to a period of hundreds of milliseconds. The authors achieved this by protecting the stored state after it was produced. Overall, the combination of the two ingredients enabled entanglement to be generated almost ten times faster than it was lost. Thanks to this achievement, Humphreys and colleagues demonstrated a deterministic delivery of

entanglement roughly every 100 milliseconds.

The long-awaited advances of these two research groups demonstrate that the prospect of realizing functional quantum networks — either on the local scale between superconducting modules or on a larger scale between communication nodes that are connected by optical fibres — is getting closer to reality (see *Nature* 554, 289–292; 2018). The numbers still need to be improved; in particular, the rate at which long-distance entanglement can be delivered remains too low for practical applications. However, an increase by a factor of about 100 should be achievable in the near future.

To demonstrate large-scale and long-haul quantum networks, a combination of techniques and tools will be necessary. Among other methods, complementary approaches that rely on ensemble-based quantum memories are being developed at a fast pace<sup>11</sup>. When combined with massive multiplexing in time, frequency or space — a process that has been necessary for the development of the Internet — these methods should be able to provide entanglement at a high rate.

Another important goal will be the realization of an efficient quantum converter that links microwave photons to optical photons<sup>12</sup>.

Such a device should enable the platforms used in these two studies, which have different capabilities, to be connected. Putting these blocks together will be a tremendous challenge for science and engineering, but it promises to lead to versatile networks in which quantum processors are interconnected through a quantum-communication web. ■

**Julien Lurat** is at the *Laboratoire Kastler Brossel (Sorbonne Université, CNRS, ENS, Collège de France), Sorbonne Université, Campus Pierre et Marie Curie, 75005 Paris, France.*  
e-mail: [julien.lurat@sorbonne-universite.fr](mailto:julien.lurat@sorbonne-universite.fr)

1. Kurpiers, P. *et al. Nature* **558**, 264–267 (2018).
2. Humphreys, P. C. *et al. Nature* **558**, 268–273 (2018).
3. Kimble, H. J. *Nature* **453**, 1023–1030 (2008).
4. Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).
5. Axline, C. J. *et al. Nature Phys.* <https://doi.org/10.1038/s41567-018-0115-y> (2018).
6. Reiserer, A. & Rempe, G. *Rev. Mod. Phys.* **87**, 1379–1418 (2015).
7. Hucul, D. *et al. Nature Phys.* **11**, 37–42 (2015).
8. Duan, L.-M., Lukin, M. D., Cirac, J. I. & Zoller, P. *Nature* **414**, 413–418 (2001).
9. Chou, C.-W. *et al. Science* **316**, 1316–1320 (2007).
10. Pfaff, W. *et al. Science* **345**, 532–535 (2014).
11. Maring, N. *et al. Nature* **551**, 485–488 (2017).
12. Andrews, R. W. *et al. Nature Phys.* **10**, 321–326 (2014).

#### IMMUNOLOGY

## Tumour tamed by transfer of one T cell

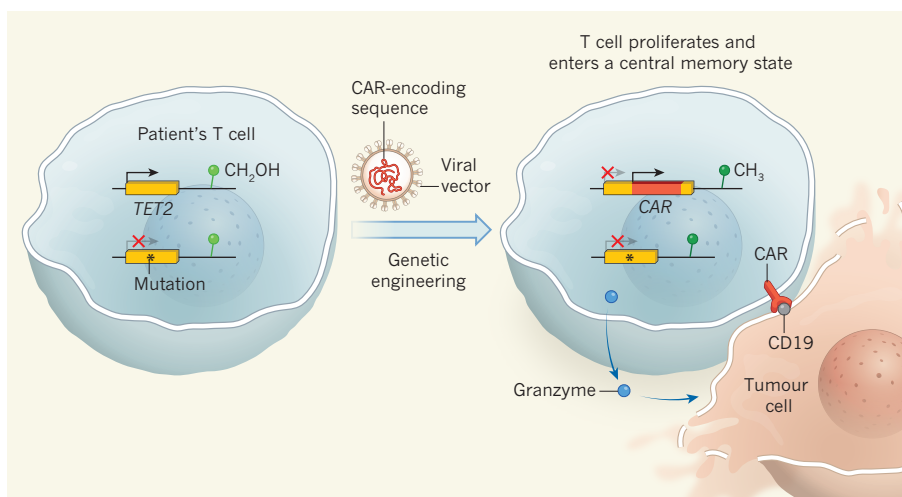
**The T cells of the immune system can be engineered to target a tumour, but why some people respond better than others to such therapy is unclear. One patient's striking response to treatment now offers some clues. [SEE LETTER P.307](#)**

MARCELA V. MAUS

**T**he use of genetically engineered immune cells to target tumours is one of the most exciting current developments in cancer treatment. In this approach, T cells are taken from a patient and modified *in vitro* by inserting an engineered version of a gene that encodes a receptor protein. The receptor, known as a chimaeric antigen receptors (CAR), directs the engineered cell, called a CAR T cell, to the patient's tumour when the cell is transferred back into the body. This therapy can be highly effective for tumours that express the protein CD19, such as B-cell acute leukaemias<sup>1,2</sup> and large-cell lymphomas<sup>3,4</sup>. However, some people do not respond to CAR T cells, and efforts to optimize this therapy are ongoing. On page 307, Fraietta *et al.*<sup>5</sup> report the fortuitous identification of a gene that positively affected one person's response to treatment with CAR T cells.

Therapies involving engineered immune cells use viral vectors based on retroviruses or lentiviruses to insert a DNA sequence, such as one encoding a CAR, into a person's T cells. However, given that there is no control over where the sequence inserts into the genome, it is possible that the engineered gene could insert at a location that disrupts another, important gene. In the early 2000s, a clinical trial<sup>6</sup> enrolled people with immunodeficiencies arising from the lack of a functional copy of a particular immune gene. The trial used viral vectors to insert a wild-type copy of this gene into their stem cells. Unfortunately, however, several people developed uncontrolled T-cell proliferation that evolved into T-cell leukaemia. This event was linked<sup>7</sup> to the gene inserting within the sequence of the *LMO2* gene, disrupting the normal regulation of *LMO2*.

The pattern of genomic integration sites for various viral vectors has been found to be specific for a given combination of vector



**Figure 1 | Tumour targeting by CAR T cells.** If a patient's T cells are engineered to express a version of an immune-cell receptor called a CAR, the cells can target tumour cells that express a specific protein, such as CD19. However, not everyone responds to this treatment. Fraietta *et al.*<sup>5</sup> report that one patient's response to CAR T-cell treatment has revealed a gene that can affect therapy success. The patient had a mutation in one of their copies of the *TET2* gene. *TET2* encodes an enzyme that converts methyl ( $\text{CH}_3$ ) groups attached to DNA into hydroxymethyl ( $\text{CH}_2\text{OH}$ ) groups. This type of change is known as an epigenetic modification. When a CAR-encoding sequence was introduced into the patient's T cells using a viral vector, in one cell the CAR sequence inserted into the patient's non-mutated copy of *TET2* and disrupted the gene, thereby generating a cell that lacked any functional copies of *TET2*. The clonal descendants of this cell eradicated the patient's tumour. The lack of *TET2* altered the cell's profile of epigenetic modifications, which can affect gene expression. This *TET2* deficiency was associated with an increase in the expression of tumour-killing factors such as the enzyme granzyme, as well as entry into a cellular state called the central memory state, which stops the cells from entering a dysfunctional mode called exhaustion.

and cell type<sup>8</sup>. In a study of people who had T cells modified using retroviral vectors, the integration events were not implicated as the cause of any cancers<sup>9</sup>. Lentiviral vectors integrate randomly into the genome but tend to preferentially locate at sites of transcriptionally active genes<sup>10</sup>. Although random integration is generally thought to be safe, any disruption of the genome nevertheless confers a risk of unwanted consequences.

The effectiveness of treatments involving CAR T cells has been linked to the persistence and proliferation of the CAR T cells in the person's body, and this can be affected by factors including the disease subtype, the molecular design of the CAR used, and even the manufacturing process<sup>1</sup>. Fraietta *et al.* report the unusual response of a person in a clinical trial whose CAR T cells targeted a CD19-expressing tumour called chronic lymphocytic leukaemia. In this case, disruption of the gene into which the CAR sequence had been inserted had a direct and beneficial effect on the clinical outcome.

The patient began to show a noticeable response to treatment two months after receiving a second dose of CAR T cells. Tumour regression normally occurs within a month if treatment is successful, so the authors investigated the reason for the delay in this case. Crucially, they analysed the nature of the CAR T cells at peak concentrations in the blood during tumour regression. Fraietta and colleagues made the surprising observation that these CAR T cells consisted almost exclusively of a clonal

population descended from a single cell.

This single cell's progeny divided over time until the cellular descendants reached a tipping point that eliminated the entire tumour. It is remarkable that the minimally effective and curative dose of this form of immunotherapy can be the introduction of just one cell. This raised the question of why introducing the CAR sequence to this specific T cell caused such an effective antitumour response.

In this clonal population of T cells, the CAR sequence had inserted into a copy of the *TET2* gene, preventing the gene from encoding a functional protein. The patient's other copy of *TET2* had a mutation, so insertion of the CAR sequence generated T cells that lacked *TET2* protein. *TET2* is an enzyme, also called methylcytosine dioxygenase, that catalyses a hydroxylation reaction that alters methyl groups attached to DNA (Fig. 1). Such modifications of DNA or its associated proteins are known as epigenetic modifications, and they can affect gene expression in some cases. When Fraietta and colleagues compared the patient's T cells that lacked the CAR insertion with those into which the CAR had been inserted, the overall epigenetic profile was similar. However, differences in the structure of the DNA-protein complex called chromatin were observed in genes involved in T-cell function, including *CD28*, *ICOS* and the gene that encodes interferon- $\gamma$ .

*TET2* mutations have previously been associated with clonal blood-cell alterations linked to a risk of disease or blood cancers (a phenomenon known as clonal

haematopoiesis)<sup>11</sup>. However, the patient's T cells that lacked *TET2* did not give rise to either aberrant T-cell proliferation or cancer. After tumour elimination, the number of CAR T cells decreased appropriately, replicating the normal pattern for a T-cell population (increasing in response to its target and declining after target elimination).

The authors used genetic engineering to remove *TET2* in human T cells *in vitro*. Analysis of these cells revealed a connection between the absence of *TET2* and the promotion and maintenance of T cells in a cellular state known as a central memory state. This state helps to prevent the cells from entering a dysfunctional mode called exhaustion, which is linked to ineffective tumour targeting by T cells. The absence of *TET2* was also linked to an increase in long-term T-cell memory.

Fraietta *et al.* observed that human T cells lacking *TET2* made fewer immune signalling molecules called cytokines than did cells that had *TET2*. Disruption of *TET2* was also linked to an increase in the level of the enzymes perforin and granzyme, which are components of the tumour-killing machinery of T cells. These roles of *TET2* in T-cell function and memory were previously unknown.

These remarkable findings might suggest that targeting *TET2* in human T cells through drug-mediated inhibition or gene-editing techniques could increase the effectiveness of CAR T-cell treatment for other patients. If so, perhaps the dose of CAR T cells needed might be only a few cells, rather than the usual 50 million to 500 million cells. This would shorten the waiting time for CAR T cells and lower the substantial manufacturing costs. However, given the known associations of *TET2* mutations with certain disease states, this approach might run the risk of generating a malignancy.

Enhancing CAR T-cell function is an area of active research, and other options to achieve this goal have been proposed. Inserting a CAR sequence at the genomic location where the natural version of the gene resides enhances the activity and persistence of CAR T cells in a mouse model<sup>12</sup>. Other groups have reported progress<sup>13</sup> in making CAR T cells resistant to inhibitory checkpoint-signalling pathways that hinder T-cell function.

Will one of the many possible approaches be preferable to the others? Unfortunately, animal studies are not always predictive of results in humans, so clinical trials are the only way to answer this definitively. The good news is that it seems likely that many of these approaches will enhance efficacy and safety, so there is hope that the use of CAR T cells to treat cancer will become even more successful in the years to come. ■

**Marcela V. Maus** is in the Department of Medicine, Division of Hematology and Oncology, Harvard Medical School, and the Massachusetts General Hospital Cancer



Center, Boston, Massachusetts 02114, USA.  
e-mail: mvmaus@mg.harvard.edu

1. Maus, M. V. & June, C. H. *Clin. Cancer Res.* **22**, 1875–1884 (2016).
2. Park, J. H. *et al. N. Engl. J. Med.* **378**, 449–459 (2018).
3. Neelapu, S. S. *et al. N. Engl. J. Med.* **377**, 2531–2544 (2017).

4. Schuster, S. J. *et al. N. Engl. J. Med.* **377**, 2545–2554 (2017).
5. Fraietta, J. A. *et al. Nature* **558**, 307–312 (2018).
6. Hacein-Bey-Abina, S. *et al. N. Engl. J. Med.* **346**, 1185–1193 (2002).
7. Hacein-Bey-Abina, S. *et al. Science* **302**, 415–419 (2003).
8. Biasco, L. *et al. EMBO Mol. Med.* **3**, 89–101 (2011).

9. Scholler, J. *et al. Sci. Transl. Med.* **4**, 132ra53 (2012).
10. Schröder, A. R. W. *et al. Cell* **110**, 521–529 (2002).
11. Buscarlet, M. *et al. Blood* **130**, 753–762 (2017).
12. Eyquem, J. *et al. Nature* **543**, 113–117 (2017).
13. Ren, J. *et al. Clin. Cancer Res.* **23**, 2255–2266 (2016).

This article was published online on 30 May 2018.

## BIOPHYSICS

# Remote wiggling helps cold enzymes work

**Mutations introduced far from the active site of an enzyme can cause local unfolding that increases enzyme activity. This finding suggests how organisms that live in the cold can speed up biochemical reactions. [SEE LETTER P.324](#)**

ASHOK A. DENIZ

The biochemical sciences have tended to focus on processes that take place at ‘physiological’ temperatures of around 37 °C. But much of Earth’s surface is covered with ocean, ice or snow, and is replete with organisms that function at much lower temperatures. Life in these environments requires suitable biological adaptations, for example in the enzymes that maintain the chemical environment of the cell. On page 324, Saavedra *et al.*<sup>1</sup> shed light on a biophysical mechanism for such low-temperature adaptation that operates at the molecular level. Their results show strikingly that protein modifications distant from an enzyme’s active site can modulate localized unfolding of the enzyme — effectively, wiggling of parts of the enzyme’s structure — that can control several facets of enzyme-reaction mechanisms.

Physical chemists have long known that the rate of chemical reactions depends on temperature, and that reaction rates generally decrease as temperatures drop. This temperature dependence also applies to enzyme-catalysed reactions, raising the intriguing question of how psychrophilic organisms (which live at low temperatures) can maintain their repertoire of enzyme-mediated functions. Related enzymes in psychrophilic organisms and in mesophilic organisms (which live at physiological temperatures) have similar activities — that is, the reactions they catalyse occur at similar rates<sup>2</sup>. For this to occur, the functional parameters of the cold-adapted enzymes must have been tuned to compensate for the lower temperatures.

A clue to how this tuning could occur came from previous observations<sup>3</sup> that psychrophilic enzymes tend to have more surface glycine mutations (in which an amino-acid residue on the protein’s surface is replaced by a glycine residue) far from the active site than do similar

enzymes in mesophilic organisms. However, the mechanistic details of this phenomenon were poorly understood. Saavedra *et al.* used an enzyme called adenylate kinase to test the mechanism by which such glycine mutations act from a distance to alter enzyme function.

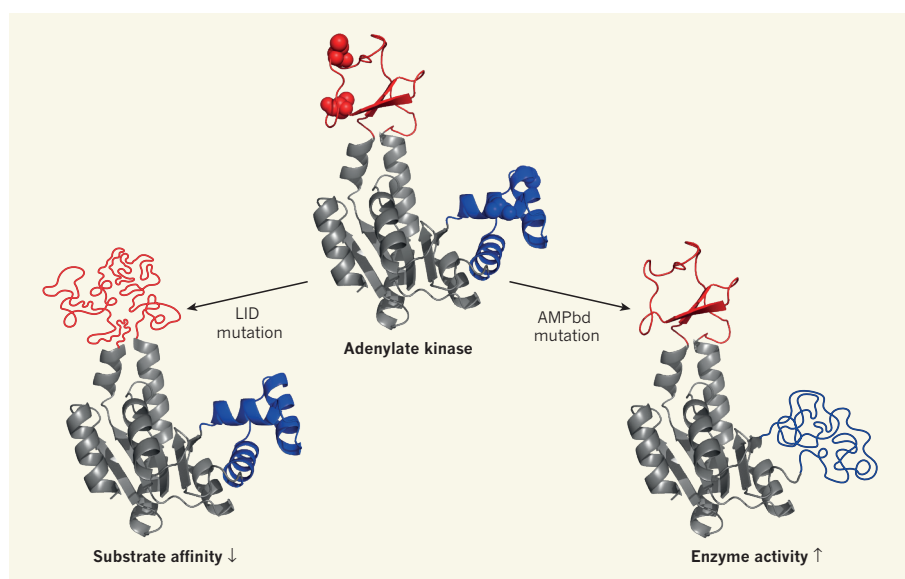
Adenylate kinase catalyses reactions that help to maintain a balance of adenosine phosphates (molecules that act as the energy currency of cells). The authors chose this mesophilic enzyme because it has been used extensively as a model system for investigating enzyme biophysics, biochemistry and folding, including by researchers from the same

laboratory as Saavedra and co-workers.

In the present work, the authors tested the previously discussed idea<sup>2</sup> that tuning of entropy — a measure of disorder — is a major driving force in the adaptation of enzymes to low temperatures. They sought to probe the effect of surface glycine mutations, at locations far from the active site, that might change the ‘wiggling’ (an entropic effect) of the protein without changing its overall folded structure.

There are three domains in adenylate kinase: the CORE domain, which contains much of the active site, and LID and AMPbd, each of which contains part of the active site. The authors studied glycine mutations in both the LID and the AMPbd domains by using a combination of biophysical and structural techniques. These studies included measuring the stability of the enzyme variants and their binding affinity to a mimic of the enzyme’s substrate, and a more detailed characterization of the protein states and structural fluctuations by using nuclear magnetic resonance spectroscopy.

Taken together, Saavedra and colleagues’ results demonstrate that adenylate kinase exists in at least three different states, and that the mutations change the relative occupancy (stability) of these states. Compared with the



**Figure 1 | Entropic tuning of enzyme function allows cold adaptation.** Saavedra *et al.*<sup>1</sup> made mutants of the enzyme adenylate kinase, replacing non-glycine amino-acid residues with glycine residues at the surface of either the LID domain (red) or the AMPbd domain (blue). Both types of mutation caused local unfolding of the enzyme, increasing its disorder (entropy), and altered the enzyme’s functional behaviour, despite being distant from the active site. The LID mutations decreased the affinity of adenylate kinase for its substrates, whereas the AMPbd mutations increased the enzyme activity. Such entropic tuning of function might be an evolutionary mechanism that allows enzymes to cope with low temperatures, which usually slow enzymatic reactions.

wild-type protein, both the LID and the AMPbd mutants decrease the occupancy of the fully folded structure, but they increase the stability of two different states in which either the LID or the AMPbd domain is locally unfolded (Fig. 1). The increased stability of these locally unfolded states stems from the fact that the footprint of a glycine amino-acid residue is smaller than that of other amino-acid residues, which means that the protein chains in the glycine mutants are more flexible than those in the wild-type protein. Remarkably, these two types of local unfolding alter different aspects of the enzyme's function: LID unfolding decreases its binding affinity, whereas AMPbd unfolding increases its activity.

A particularly interesting aspect of the authors' work is that it helps us to understand how surface glycine mutations act at a distance to support cold adaptation. This phenomenon might seem mysterious at first glance, but it is encapsulated in the idea of allosteric regulation<sup>4,5</sup> — a common form of enzyme regulation in which the binding of a molecular partner at a site distant from the active site

affects enzyme activity. The conventional view of allosteric regulation has been that binding of the partner causes small structural changes that propagate through the protein to alter the structure of the active site. However, there is now evidence for mechanisms involving changes in the dynamics (entropy), rather than in the structure, of the unbound state for many cases of allosteric regulation<sup>6–8</sup>. The current work provides striking tests and examples of such entropic allosteric regulation for different enzyme properties.

Saavedra and colleagues' results have substantial implications for the evolution of enzyme function. However, their mechanistic proposal for cold adaptation was tested for just one model enzyme, so its relevance for other cold-adapted enzymes requires further testing. Deeper insight into the biophysical mechanisms of cold adaptation will also benefit from more-detailed views of the structural fluctuations of enzymes afforded by single-molecule experiments<sup>9,10</sup>. Nevertheless, the findings open up new avenues for exploring allosteric control of multiple modes of protein function,

both in natural evolution and in rational protein engineering<sup>11</sup> for biotechnology. ■

**Ashok A. Deniz** is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California 92037, USA.  
e-mail: deniz@scripps.edu

1. Saavedra, H. G., Wrabl, J. O., Anderson, J. A., Li, J. & Hilser, V. J. *Nature* **558**, 324–328 (2018).
2. Siddiqui, K. S. & Cavicchioli, R. *Annu. Rev. Biochem.* **75**, 403–433 (2006).
3. Fields, P. A. & Somero, G. N. *Proc. Natl Acad. Sci. USA* **95**, 11476–11481 (1998).
4. Monod, J., Wyman, J. & Changeux, J. P. *J. Mol. Biol.* **12**, 88–118 (1965).
5. Koshland, D. E. Jr, Némethy, G. & Filmer, D. *Biochemistry* **5**, 365–385 (1966).
6. Cooper, A. & Dryden, D. T. F. *Eur. Biophys. J.* **11**, 103–109 (1984).
7. Wand, A. J. *Curr. Opin. Struct. Biol.* **23**, 75–81 (2013).
8. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. *Nature* **508**, 331–339 (2014).
9. Ferreon, A. C. M., Ferreon, J. C., Wright, P. E. & Deniz, A. A. *Nature* **498**, 390–394 (2013).
10. Deniz, A. A., Mukhopadhyay, S. & Lemke, E. A. *J. R. Soc. Interface* **5**, 15–45 (2008).
11. Dokholyan, N. V. *Chem. Rev.* **116**, 6463–6487 (2016).

This article was published online on 6 June 2018.

1% chance in any given year); such flooding events can be severe (Fig. 1). To do this, they used a database of the world's coastline characteristics<sup>5</sup>, which includes land-elevation data measured by radar. The authors combined the data with their scenarios of sea-level rises, and found that an area of 540,000 square kilometres is already at risk of 1-in-100-year coastal flooding events. For the scenario in which the global temperature rise is mitigated to 1.5 °C, this area increases to 620,000 km<sup>2</sup> by 2100, and to 702,000 km<sup>2</sup> by 2300 (values correspond to the 50th percentile of the range of predicted values for sea-level rises). In the absence of mitigation, they find that the area at risk by 2100 (708,000 km<sup>2</sup>) is not much different from that in the mitigation scenario, but increases to 1,630,000 km<sup>2</sup> by 2300, which is three times the area at risk today.

In the third component of Brown and colleagues' study, the authors consider the number of people at risk from coastal flooding. If global warming is kept to 1.5 °C, they find that 1.5–2.1% of the global population will be exposed to a 1-in-100-year coastal flooding by 2100, compared with 4.3–5.4% of the global population in the non-mitigation scenario. In other words, more than half of the potential population exposure can be avoided by 2100 if global warming is capped. It is important to keep in mind, however, that population exposure does not depend only on the amount of sea-level rise — the number of people exposed could decrease if people move away from the coast, for example.

Given that sea levels will rise irrespective of future global temperature changes, Brown *et al.* stress that at least some action will need to be taken to adapt. However, their calculations of the land area exposed to sea-level rise do

## CLIMATE CHANGE

# How humans and rising seas affect each other

**The Paris climate agreement aims to limit global warming to no more than 2 °C. An analysis suggests that this will greatly reduce the risks of sea-level rise for coastal communities, but that it will take time to see the benefit.**

AIMÉE SLANGEN

Coastal zones are among the most densely populated areas in the world<sup>1</sup>, but they are threatened by rising sea levels caused by climate change. Writing in *Earth's Future*, Brown *et al.*<sup>2</sup> estimate the impact of sea-level rise in terms of the land area and the number of people exposed, for several scenarios in which global warming is limited to different temperature increases. Their study shows that the amount of exposure to sea-level rise depends on our ability to cap global temperature changes, and that the main benefits of this cap will be seen only after 2100.

Brown and colleagues began their investigation by simulating how global temperature will change in response to different scenarios of greenhouse-gas emissions, using a simple computational model of the Earth system<sup>3</sup>. They then calculated the global mean sea-level rise that would occur as a result of the projected temperature changes, assuming that the main causes of sea-level rise are the expansion of the volume of ocean water associated

with warming, and the melting of land-based ice (that is, ice over land in glaciers, Antarctica and Greenland). To account for the fact that sea-level rise does not occur uniformly across Earth, they then scaled their time series of global mean sea levels with previously reported projections<sup>4</sup> of regional patterns of sea-level change for 2100.

In the scenario in which global temperature increases are capped at 1.5 °C before 2100, they find a median sea-level rise of 0.4 metres by 2100 and of 1 m by 2300. By contrast, in the scenario in which temperatures continue to increase as they are doing now, sea-level rises are 0.8 m in 2100 and a staggering 4.5 m in 2300. These results show that there will be some sea-level rise regardless of efforts to mitigate climate change, because sea levels will not immediately stop rising when the temperature targets are met. However, the effect of capping global temperatures early will be increasingly felt after 2100 and lead to significantly less sea-level rise by 2300.

In a second step, Brown *et al.* considered the area of land that has an average chance of being flooded once every 100 years (that is, a





**Figure 1 | Flooding in New Orleans, Louisiana, after Hurricane Katrina, 2005.** Brown *et al.*<sup>2</sup> report that flooding events, such as that caused by Hurricane Katrina, will affect more land and people in the future as a result of sea-level rises associated with global warming.

not consider the effects of coastal-protection measures, such as the construction of dykes or dunes. If such measures were considered in the analysis, then the area and number of people exposed would change. Moreover, the construction of coastal defences will be largely driven by economic considerations, which will be different for different countries.

The authors define land at risk from sea-level rise as the 1-in-100-year coastal flood-plain. But will people be driven away from such land by the infrequent floods, or will they accept the occasional inundation, moving away only temporarily as needed? It might be better to consider the area of land that will become permanently flooded to make a more-direct estimate of the population that will be exposed to sea-level rise.

A limitation of the study is that Greenland, Antarctica and all glaciers elsewhere are lumped together as the land ice that contributes to sea-level rise. However, ice will probably melt at different rates in each of these regions. This will cause the relative contributions of the different sources to change over time. One way to improve the regional estimates of sea-level rise would therefore be to scale the contributions of the ice sheets according to their individual effects.

Nevertheless, studies such as those of Brown and colleagues are essential, because they show the complexity of the climate system's response to change and how this affects society. By directly connecting the effects of climate change to the consequences for humans, the authors clearly show that climate mitigation needs to happen now for a better future. ■

**Aimée Slangen** is in the Department of Estuarine and Delta Systems, NIOZ Royal Netherlands Institute for Sea Research and Utrecht University, 4401 NT Yerseke, the Netherlands.

e-mail: [aimée.slangen@nioz.nl](mailto:aimée.slangen@nioz.nl)

1. McGranahan, G., Balk, D. & Anderson, B. *Environ. Urban.* **19**, 17–37 (2007).
2. Brown, S. *et al.* *Earth's Future* **6**, 583–600 (2018).
3. Goodwin, P. *Clim. Dyn.* **47**, 2219–2233 (2016).
4. Slangen, A. B. A. *et al.* *Clim. Change* **124**, 317–332 (2014).
5. Vafeidis, A. T. *et al.* *J. Coast. Res.* **24**, 917–924 (2008).

#### GENE REGULATION

## A new phase in transcription

**A subunit of the enzymatic complex P-TEFb can induce compartmentalization of proteins into liquid-like droplets in cells. This phase separation might help P-TEFb to promote gene transcription. [SEE LETTER P.318](#)**

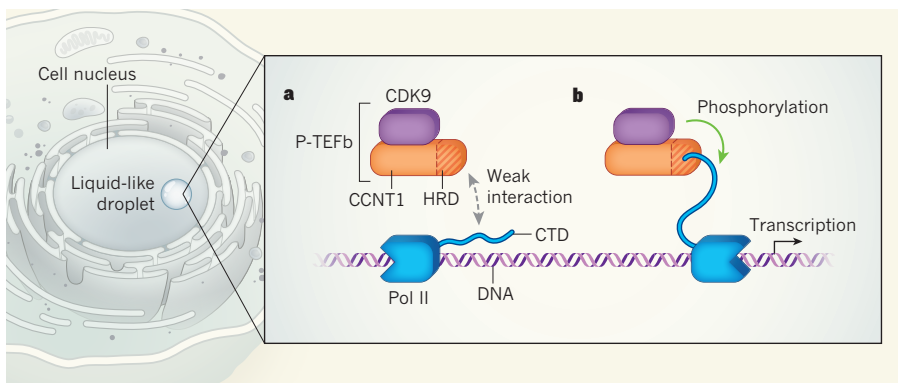
JAMES A. GOODRICH & DYLAN J. TAATJES

**G**enomic DNA is not indiscriminately transcribed into RNA. Instead, select sequences are transcribed at any given time or in any given cell type. How this process is regulated has been an active area of research for decades, because of its complexity and importance in embryonic development and disease. On page 318, Lu *et al.*<sup>1</sup> describe an aspect of transcription regulation that involves liquid-liquid phase separation — a process by which proteins, nucleic acids and other molecules self-organize into liquid-like droplets to enable subcellular compartmentalization.

Transcription is catalysed by the enzyme

RNA polymerase II (Pol II). The activity of Pol II is, in turn, regulated by a complex called P-TEFb, which consists of the protein cyclin T1 (CCNT1) and the kinase enzyme CDK9. P-TEFb binds to the carboxy-terminal domain (CTD) of Pol II (ref. 2) — this domain consists of 52 repeats of 7 amino acids, 5 of which can be phosphorylated. The CCNT1 subunit regulates the activity of CDK9, which can phosphorylate the Pol II CTD many times to help control Pol II function<sup>3</sup>.

Lu *et al.* first investigated which region of CCNT1 is responsible for regulating CDK9 activity. They found that mutations in a domain of CCNT1 rich in the amino acid histidine led to defects in the activation of



**Figure 1 | Phase separation in transcription regulation.** **a**, The P-TEFb complex consists of two proteins, CCNT1 and CDK9. Lu *et al.*<sup>1</sup> have identified a histidine-rich domain (HRD) in CCNT1 that promotes liquid–liquid phase separation — a process in which P-TEFb and the enzyme RNA polymerase II (Pol II) coalesce inside a liquid-like droplet in the cell nucleus, within which they weakly interact. **b**, The authors provide evidence that these weak interactions within phase-separated droplets enhance direct, functional interactions between P-TEFb and Pol II that allow CDK9 to phosphorylate the carboxy-terminal domain (CTD) of Pol II, promoting the ability of Pol II to transcribe DNA.

Pol II-mediated transcription *in vitro*. Notably, removal of this histidine-rich domain (HRD) decreased phosphorylation of the full-length Pol II CTD by P-TEFb, but had no effect on phosphorylation of a truncated, nine-repeat version of the CTD. The Pol II CTD is known to undergo phase separation into liquid-like droplets<sup>4</sup>, but this ability declines as the length of the CTD decreases<sup>5</sup>. The authors' results therefore hinted that the HRD might be involved in phase separation, and that this process might improve the ability of P-TEFb to phosphorylate the full-length Pol II CTD.

The researchers then fused CCNT1 to a DNA-binding domain that directed it to sites at which transcription begins. This fusion strategy allowed the group to evaluate the function of CCNT1 on its own, away from other proteins that would typically guide it to DNA. The fusion protein activated transcription, but activation decreased if HRD was deleted. This finding further supports the idea of CCNT1 being involved in phase separation, because other proteins that induce phase separation behave similarly if fused to DNA-binding domains<sup>4</sup>. Presumably, Pol II is recruited into phase-separated droplets formed by CCNT1, thus concentrating Pol II around DNA and promoting the activation of transcription.

Next, Lu *et al.* carried out a series of experiments that demonstrated that the HRD of CCNT1 does indeed promote phase separation *in vitro* and in cells. As expected for a domain that serves a key regulatory function, the CCNT1 HRD is evolutionarily conserved among vertebrates. Moreover, the authors found that a similar domain in a kinase called DYRK1A also regulated that enzyme's activity and promoted phase separation.

Finally, the group investigated whether the HRD influences the location and dynamics of CCNT1 in the nuclei of living cells. Using two complementary techniques to track CCNT1 movement, the researchers provided evidence that the presence of the HRD facilitates the

retention of CCNT1 at actively transcribing genes. Consistent with this finding, CCNT1 formed concentrated clusters called speckles in the cell nucleus in an HRD-dependent manner. These speckles were dispersed by 1,6-hexanediol — a hydrophobic compound that disrupts phase separation.

Lu and colleagues' work suggests that the P-TEFb complex can trigger the formation of liquid-like droplets around transcriptionally active genomic regions. However, P-TEFb does not require phase separation to interact with the Pol II CTD (ref. 2), and so it will be challenging to determine the exact contribution of phase separation to P-TEFb's function in transcriptional regulation. The authors show that the HRD augments the ability of CCNT1 to directly interact with the Pol II CTD. This observation can at least partially explain the defects in transcription caused by HRD mutations, but changes in phase separation are probably also involved.

The authors propose that compartmentalization through phase separation, which involves multiple weak interactions between proteins within a droplet, might concentrate P-TEFb and Pol II together. This, in turn, probably promotes functional interactions between P-TEFb and Pol II, leading to highly efficient phosphorylation of the CTD (Fig. 1). In this way, phase separation could ensure robust regulation of gene-expression programs, and enforce high-level activation of genes to enable rapid or sustained responses to extracellular stimuli.

Lu and co-workers' study raises many interesting questions. For example, do P-TEFb complexes selectively associate only with nuclear phase-separated domains that contain Pol II? Or can P-TEFb also associate with phase-separated droplets that are devoid of Pol II? A transcriptional repressor protein called HP1 $\alpha$  can also promote phase separation<sup>6,7</sup>. Thus, phase-separated domains seem to demarcate both transcriptionally active and transcriptionally inactive genomic regions. It

remains to be seen whether the HP1 $\alpha$  and the P-TEFb compartments are biophysically distinct, and whether each can selectively exclude specific sets of proteins, nucleic acids or metabolite molecules to effectively repress or promote transcription, respectively.

The impact of CCNT1 HRD mutations on global gene-expression patterns also remains to be defined. On the basis of current models<sup>8</sup>, it seems plausible that highly expressed genes will be most sensitive to HRD mutations, because their elevated expression might require high local concentrations of Pol II and transcription factors in phase-separated droplets. If this is indeed the case, cells that carry HRD mutations might be less able than wild-type cells to mount effective transcriptional responses to DNA damage or viral infection, or might fail to maintain cell-type-specific gene-expression programs over time. Such defects could contribute to cancer and other diseases. To our knowledge, any clinical manifestations of HRD mutations are unknown, and await discovery.

P-TEFb phosphorylates other transcriptional regulatory proteins in addition to Pol II (ref. 9). As such, inhibition of P-TEFb kinase activity shows promise as a therapeutic strategy for combating a variety of diseases, including HIV infection and cancer<sup>10</sup>. Compounds that inhibit P-TEFb repress transcription across the entire genome<sup>3</sup>, and it has been assumed that the inhibitors alter the function of P-TEFb's numerous targets for phosphorylation<sup>9</sup>. Is phase separation also involved? Phase-separated domains can reversibly form and dissolve, and this behaviour is sensitive to phosphorylation<sup>11</sup>. Notably, Pol II CTD phosphorylation seems to either promote or prevent phase separation, depending on which other proteins are present in the droplet<sup>1,4,5</sup>. Inhibition of P-TEFb kinase activity might therefore alter the formation and dissolution of phase-separated droplets. This intriguing possibility remains to be rigorously tested. ■

**James A. Goodrich and Dylan J. Taatjes**  
are in the Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80303, USA.  
e-mails: james.goodrich@colorado.edu;  
dylan.taates@colorado.edu

1. Lu, H. *et al.* *Nature* **558**, 318–323 (2018).
2. Ebmeier, C. C. *et al.* *Cell Rep.* **20**, 1173–1186 (2017).
3. Zhou, Q., Li, T. & Price, D. H. *Annu. Rev. Biochem.* **81**, 119–143 (2012).
4. Kwon, I. *et al.* *Cell* **155**, 1049–1060 (2013).
5. Boehning, M. *et al.* Preprint at bioRxiv <http://dx.doi.org/10.1101/316372> (2018).
6. Larson, A. G. *et al.* *Nature* **547**, 236–240 (2017).
7. Strom, A. R. *et al.* *Nature* **547**, 241–245 (2017).
8. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. *Cell* **169**, 13–23 (2017).
9. Sanso, M. *et al.* *Genes Dev.* **30**, 117–131 (2016).
10. Ferguson, F. M. & Gray, N. S. *Nature Rev. Drug Disc.* **17**, 353–377 (2018).
11. Li, P. *et al.* *Nature* **483**, 336–340 (2012).

This article was published online on 30 May 2018.





**Cover image**  
Camille Seaman

**Editor, Nature**  
Philip Campbell

**Publishing**  
Richard Hughes

**Insights Editor**  
Ursula Weiss

**Subeditors**  
Dinah Loon  
Kristen Harley,  
Chariklia Rouki

**Art Editor**  
Nik Spencer

**Sponsorship**  
Reya Silao

**Production**  
Ian Pope

**Marketing**  
Steven Hurst

**Editorial Assistant**  
Jasmine Delves

The Campus  
4 Crinan Street  
London N1 9XW, UK  
Tel: +44 (0) 20 7833 4000  
e: nature@nature.com

**SPRINGER  
NATURE**

An uninterrupted thread of discovery, triumph, and mystery—tempered by disaster, exploitation, and misery—courses through humanity’s interaction with Antarctica. The relationship is riven with conflict, and heroic interventions are often required: reminiscing about unimaginable suffering and bleak prospects, geologist and explorer Sir Raymond Priestley once said “...get down on your knees and pray for Shackleton”. Now, even though Shackleton continues to inspire, the situation is largely reversed. Rather than humans seeking salvation from Antarctica’s harsh caress, it is Antarctica that must seek protection from humanity.

This Nature Insight on Antarctica peers into these many facets of Antarctica’s past, present and possible futures. In the first Review, Ed Brook and Christo Buizert synthesize 800,000 years of climate and atmospheric history to depict Antarctica’s response to and influence on the broader climate system. Steve Rintoul then reviews the vast Southern Ocean, and shows that localized processes—such as eddy-driven mixing sourced from current interaction with bottom topography—drive the overall system.

An alarming amount (more than 50 metres) of sea-level rise lurks in the Antarctic Ice Sheet, and a first step in understanding the ice’s fate is knowing the current mass imbalance. In an Analysis, the IMBIE team brings together the lengthening remote sensing record and reveals an accelerating trend of ice loss. Then, in a Review, Andy Shepherd, Helen Amanda Fricker and Sinead Louise Farrell discuss the tightly interlinked processes governing recent trends in Antarctic sea ice, ice shelves and grounded ice.

Finally, in a Perspective arising from a 2014 panel discussion attended by winners of the Tinker-Muse Prize for Science and Policy in Antarctica, Steve Rintoul and colleagues present two visions of Antarctica in 2070: in one, humanity’s appetite for resources and fossil fuels continues to increase, and governance weakens; in the other, world development proceeds with a sharper focus on conservation, mitigation, and strong governance.

The Insight brings together much of our understanding of natural science in and around Antarctica, reveals our current impacts on it, and outlines a pathway through which the ghost of Shackleton might not need to be raised, on behalf of an entire continent.

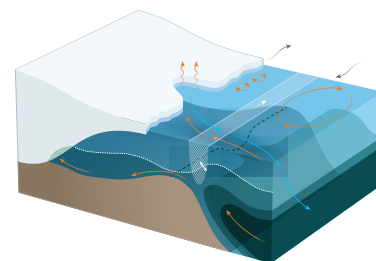
**Michael White**  
Senior Editor

### CONTENTS

#### REVIEWS

**200 Antarctic and global climate history viewed from ice cores**  
Edward J Brook & Christo Buizert

**209 The global influence of localized dynamics in the Southern Ocean**  
Stephen R Rintoul

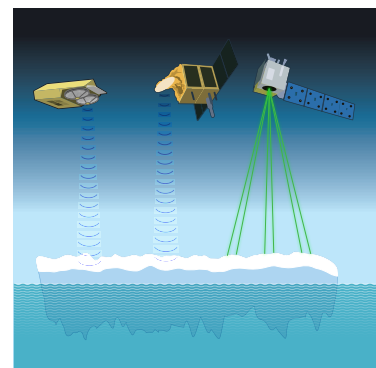


#### ANALYSIS

**219 Mass balance of the Antarctic Ice Sheet from 1992 to 2017**  
The IMBIE team

#### REVIEW

**223 Trends and connections across the Antarctic cryosphere**  
Andrew Shepherd, Helen Amanda Fricker & Sinead Louise Farrell



#### PERSPECTIVE

**233 Choosing the future of Antarctica**  
S R Rintoul, S L Chown, R M DeConto, M H England, H A Fricker, V Masson-Delmotte, T R Naish, M J Siegert & J C Xavier

# Antarctic and global climate history viewed from ice cores

Edward J. Brook<sup>1\*</sup> & Christo Buizert<sup>1</sup>

**A growing network of ice cores reveals the past 800,000 years of Antarctic climate and atmospheric composition. The data show tight links among greenhouse gases, aerosols and global climate on many timescales, demonstrate connections between Antarctica and distant locations, and reveal the extraordinary differences between the composition of our present atmosphere and its natural range of variability as revealed in the ice core record. Further coring in extremely challenging locations is now being planned, with the goal of finding older ice and resolving the mechanisms underlying the shift of glacial cycles from 40,000-year to 100,000-year cycles about a million years ago, one of the great mysteries of climate science.**

The origins of the Antarctic ice sheet are in late Eocene–early Oligocene time, about 34 million years ago; the last major phase of growth began at about 14 million years BP (before present)<sup>1,2</sup>. Coring in the present ice sheet (Box 1) has recovered records up to 800,000 years (800 kyr) old and there are hopes of extending the continuous record further. New studies on the ice sheet margin have provided unique but discontinuous samples of older ice<sup>3</sup>. Owing to the flow of ice from the interior to the ice sheet margin, ice deposited during the earliest history of the ice sheet is not likely to have been preserved, but the maximum age of extant ice is unknown.

The Antarctic ice sheet preserves a history that links Antarctica to the rest of Earth. Richly detailed and uniquely preserving the composition of the atmosphere, the record in the ice underpins much of global climate change research. Drilling and recovering long ice cores (Box 1) require specialized engineering owing to the low temperatures and the high pressures inside the ice sheet, the need to preserve the ice for analysis, and the remote locations of deep-drilling sites. The stratigraphically continuous dataset from cores in the dry ice sheet interior that extends to 800 kyr is complemented by more detailed, but younger, datasets from coastal sites with higher snowfall rates (Box 1). Collectively, Antarctic ice cores provide fundamental insights into the nature of global climate cycles driven by orbital variations, internal climate variability on sub-orbital timescales, changes in global biogeochemical cycles, Antarctic climate dynamics, abrupt climate change, and a host of other topics.

One of the strengths of the ice core record is that a wide variety of parameters, reflecting different aspects of the Earth system, can be measured in great detail in the same core. The isotopic composition of the ice is a proxy for local temperature<sup>4</sup>, while the chemical composition records the input of dust, sea salt, volcanic material, pollutants, other aerosol material, and even extraterrestrial dust<sup>5–8</sup>. Past snowfall rate, a fundamental climate parameter, can be derived from layer counting and other age constraints<sup>9</sup>. The temperature of the ice sheet itself retains a memory of past climate and can be measured in the ice core borehole<sup>10</sup>. Gradual compaction of the firn converts snow to ice, trapping small samples of the atmosphere in a matrix that is remarkably resistant to gas loss. The trapped air provides the only direct record of changes in atmospheric composition prior to modern atmospheric measurements.

Over the last decade, drilling and analysis of Antarctic ice cores have uncovered a large amount of new information. This paper reviews the myriad ways in which these data show how connected Antarctica is to

the rest of the world. Deep ice cores from the Antarctic interior show that on long timescales Antarctic climate closely follows variations in solar insolation that drive climate change globally and that it exhibits major temperature changes at the termination of ice ages. New results from high-resolution cores in high-accumulation regions provide unprecedented detail about the millennial-scale climate ‘seesaw’ between Antarctica and the Northern Hemisphere, a signature of variations in ocean heat transport related to shifts in Atlantic Ocean circulation. The dust content of ice cores reveals enhancements in dust flux in cold climates, with possible contributions to ocean productivity. Greenhouse gas data from trapped air show how global biogeochemical feedbacks contribute to climate change on long and short timescales, and that climate and greenhouse forcing are extremely tightly coupled (Box 2).

## The long view: ice age cycles

Two deep ice cores from low-accumulation regions of the East Antarctic Plateau, one at Dome Concordia<sup>11</sup> (Dome C) and a second at Dome Fuji<sup>12</sup> (Dome F, or Valkyrie Dome) provide now-iconic records of multiple glacial cycles as far back as 800 kyr and 720 kyr BP, respectively (Fig. 1). These results complement the pioneering 420-kyr dataset from the Vostok ice core<sup>13</sup>. Viewed broadly, almost all climate-related parameters in these records show major, synchronous variations across the glacial cycles, which have an average duration of about 100 kyr. The Antarctic records closely mirror other global environmental proxies, most notably the oxygen isotopic composition of benthic foraminifera in deep-ocean sediments (Fig. 1g), which is widely used as an index for global ice volume and glacial–interglacial conditions<sup>14</sup>. All Antarctic records show a characteristic sawtooth pattern on these timescales, with a gradual cooling trend from glacial inception to peak glacial conditions, followed by a relatively fast glacial termination.

The stable water isotope ratios ( $\delta D$  or  $\delta^{18}O$ ) traditionally used as temperature proxies suggest that the amplitude of East Antarctic temperature change over the glacial cycles ranges from about 6 °C to 13 °C, with the largest-amplitude oscillations<sup>11</sup> during the last 450 kyr. Ice isotope records are an indirect measure of temperature, however, and are influenced by other phenomena related to moisture transport and deposition. The ice sheet has a long thermal memory, and in locations with thick ice and high accumulation rate borehole thermometry can be used to circumvent these problems and estimate glacial–interglacial temperature change. Borehole-based temperature

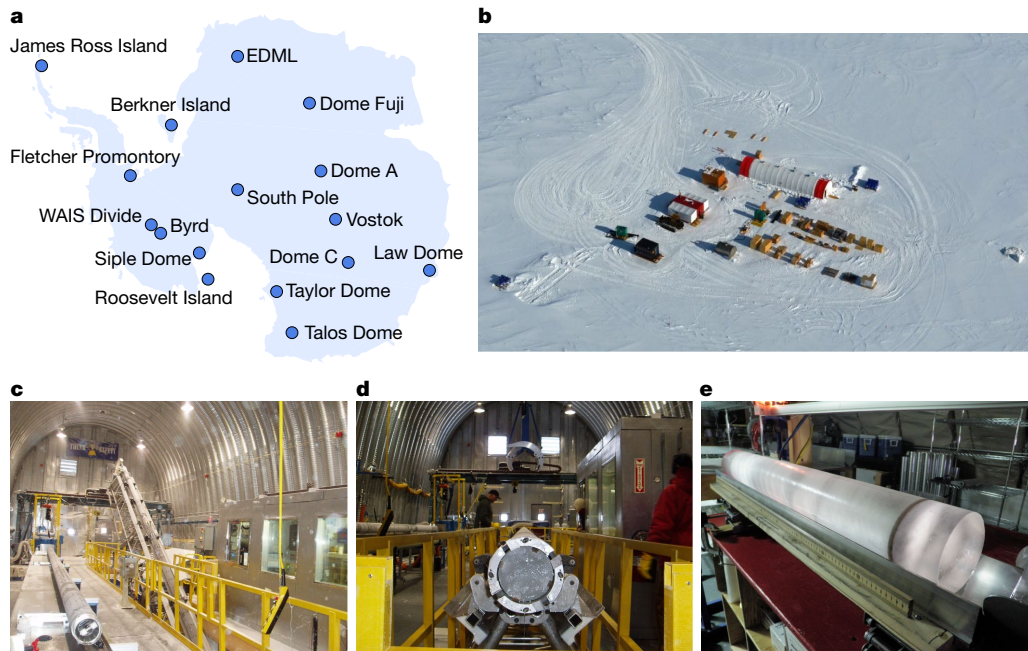
<sup>1</sup>College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA. \*e-mail: [brooke@geo.oregonstate.edu](mailto:brooke@geo.oregonstate.edu)



# Box I

## Ice core drilling

The science of ice core drilling originated in the 1950s and the first deep core in Antarctica was at Byrd Station in 1968, during the International Geophysical Year<sup>111</sup>. Since then, numerous deep-coring projects have been completed (blue circles in **a** in the figure). The longest core, at Vostok Station, reaches 3,700 m below the ice surface. The oldest so far, the EPICA Dome C ice core, extends to 800,000 years<sup>11,112</sup>. The technology for deep ice coring has gradually evolved, with recent developments in larger volume and replicate coring drills<sup>113</sup>, and new access tools that will allow quick sampling of deep ice sections, coring within bedrock and analysis in situ<sup>101–103</sup>. National archive facilities retain samples from all deep ice cores, preserving a unique resource for the scientific community. At an ice core drilling camp at the South Pole (**b**), the long arch structure contains the drill. The tipping tower of the US Deep Ice Coring Drill is shown (**c**). An ice core section in the WAIS Divide Drill (**d**) is shown immediately after a drilling run. A core section from the WAIS Divide site (**e**) shows a visible volcanic ash layer. The core is 12 cm in diameter. Image credits: **b**, US National Science Foundation (<http://spicecore.org/photos.shtml>); photographs in **c** and **d** were taken by Jay Johnson and in **e** by Heidi Roop.



reconstructions at the WAIS Divide site in west Antarctica<sup>10</sup> indicate a glacial–interglacial temperature change of  $11.3 \pm 1.8$  °C for the last termination, consistent with estimates based on stable water isotopes alone. Globally, glacial–interglacial temperature change has been estimated<sup>15</sup> at about 3.5 °C; the higher Antarctic values confirm the theory that polar temperatures change more than tropical temperatures do (polar amplification)<sup>10</sup>.

It is understood that the glacial cycles are paced by variations in Earth's orbit<sup>16</sup>, with insolation changing due to orbital eccentricity (with an approximately 100-kyr period), axial tilt (with an approximately 41-kyr period) and precession of the equinoxes (with an approximately 19–23-kyr period). In terms of radiative forcing at the top of the atmosphere, the insolation changes by themselves are too weak to drive global temperature changes, and feedbacks such as changes in greenhouse gases and the high albedo of extensive (Northern Hemisphere) glaciation are required to explain the observed climate variations<sup>10</sup>.

Several challenges remain to our understanding of glacial–interglacial dynamics, foremost of which is the lack of a clear explanation for the apparent 100-kyr periodicity of the glacial cycle, given that the 100-kyr eccentricity cycle produces negligible variations in either seasonal or annual radiative forcing (the so-called ‘100-kyr problem’)<sup>17–19</sup>. Closer inspection shows that individual cycles are not uniform in length; accurate U/Th dating of speleothems suggests that glacial terminations are actually spaced by four to five precession cycles, supporting the theory that changes in Northern Hemisphere insolation driven by precession are the predominant driver of glacial

terminations<sup>20</sup>. Several models have been put forward to ‘predict’ which Northern Hemisphere insolation maxima lead to glacial terminations and which do not<sup>19,21</sup>. Most proposed solutions to the 100-kyr problem rely on the inertia of gradually growing Northern Hemisphere ice sheets that have the ability to survive several insolation maxima, typically with a nonlinear response of ice volume to insolation<sup>17,19,22–24</sup>.

A second challenge concerns the remarkable coherence of Antarctic temperature variability and global climate change on orbital timescales, although orbital forcing due to precession acts with opposite effect in each hemisphere<sup>25</sup>. The canonical view holds that global climate responds to summer solstice insolation at 65° N, the latitude band of the large Northern Hemisphere ice sheets. Indeed, accurate dating of the Dome Fuji ice core confirms that Antarctic orbital-scale climate change follows Northern Hemisphere insolation closely, with the largest warmings concurrent with rising Northern Hemisphere summer insolation<sup>26</sup>.

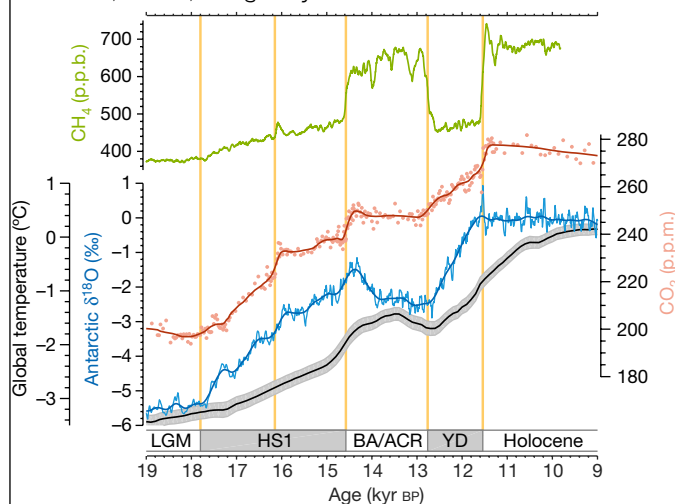
The most obvious mechanism to explain this bi-hemispheric coherence is through the globally well mixed greenhouse gases, although their link to Northern Hemisphere insolation remains incompletely understood. Rising Northern Hemisphere insolation could further drive enhanced Laurentide ice sheet meltwater runoff into the North Atlantic to reduce the Atlantic overturning circulation, which would in turn warm Antarctica via the so-called ‘bipolar seesaw’ (see section ‘The close view: millennial-scale variability’), suggesting a role for ocean circulation in synchronizing the hemispheres<sup>5,25,27</sup>. Alternatively, it has been suggested that synchronicity at precession frequencies arises from the fact that Antarctic temperature is sensitive to local summer duration (which changes in-phase with the Northern Hemisphere

## Box 2

CO<sub>2</sub> and temperature phasing during the last deglaciation

The question of the relative timing of greenhouse gas and (Antarctic) temperature change at the glacial terminations has generated substantial interest. Evaluating this phasing is complicated by the difference in age between the ice and the gas trapped inside it. Early studies suggested a substantial (600- to 1,200-year) lag of the CO<sub>2</sub> concentration rise behind Antarctic warming<sup>114–117</sup>. More recent work (see figure), based on high-accumulation cores and improved treatment of firn compaction, finds CO<sub>2</sub> and Antarctic temperature to be more or less synchronous (within uncertainty) for the last two glacial terminations<sup>84,118,119</sup>. The close relationship between Antarctic temperature and CO<sub>2</sub> obviously reflects important feedbacks and interactions between the global carbon cycle and the climate system. The early studies suggesting a long lag of CO<sub>2</sub> increase behind Antarctic warming have been misinterpreted as proof of a negligible warming effect of greenhouse gases; this is incorrect because (1) the onset of Antarctic warming is driven by changes in interhemispheric heat exchange, rather than by CO<sub>2</sub>, and (2) the global temperature rise lags CO<sub>2</sub>, rather than leads it<sup>15</sup>. However, what then does this phasing tell us? One interpretation is that the close phasing reflects the dominant role of Southern Ocean ventilation in setting atmospheric CO<sub>2</sub> levels<sup>119</sup>. Alternatively, the rise in both CO<sub>2</sub> levels and Antarctic temperature may be caused by reduced North Atlantic Deep Water formation, in which case their synchronicity reflects a common driver, rather than interdependence. Much as in Douglas Adams' *The Hitchhiker's Guide to the Galaxy*, although the answer to our question is now apparent, precisely what it signifies remains to be revealed.

The figure shows atmospheric CO<sub>2</sub> change from the WAIS Divide (Antarctica) ice core for the period 19–9 kyr ago, global temperature reconstruction<sup>15</sup>, the east Antarctic oxygen isotope stack (the water <sup>18</sup>O/<sup>16</sup>O isotope ratio anomaly relative to the present)<sup>119</sup>, and the atmospheric CH<sub>4</sub> record from the WAIS Divide ice core<sup>92</sup>. Vertical yellow bars show the timing of major inflection points in the CO<sub>2</sub> record. Grey shading around the black trace indicates uncertainty in the temperature reconstruction. LGM, Last Glacial Maximum; HS1, Heinrich Stadial 1; BA, Bølling–Allerød; ACR, Antarctic Cold Reversal; and YD, Younger Dryas.



summer solstice insolation intensity in the precession band), rather than summer peak insolation (which changes out-of-phase with the same)<sup>28</sup>. Although local insolation may play a role at some ice core

sites<sup>29</sup>, the precise timing and magnitude of temperature variations suggest that the bipolar seesaw and greenhouse gas variations are the dominant influences on orbital-scale Antarctic climate.

A third challenge is that the amplitude of the glacial cycles in Antarctic temperature, CO<sub>2</sub> and ice volume increased around 450 kyr BP, at the so-called Mid-Brunhes event, through an intensification of the interglacial climates (Fig. 1). Atmospheric CH<sub>4</sub> and other markers of tropical hydrology appear to be only weakly affected by this transition<sup>20,30</sup>. It is debated whether the Mid-Brunhes is truly a single event, or whether this apparent transition simply emerges from the orbital forcing without changes in the underlying coupling between insolation and climate<sup>19,21,31</sup>. Ocean sediments show that the current '100-kyr world' was preceded by a '41-kyr world' in which glacial cycles were paced by Earth's axial tilt<sup>14</sup>. The transition between these two occurred between 1,200 kyr and 800 kyr BP, a period not covered in the ice core record. A major goal of the international ice coring community is to recover a continuous ice core through this key transition (see section 'The future of Antarctic ice core science'), primarily to examine changes in greenhouse forcing and the temporal patterns of Antarctic climate.

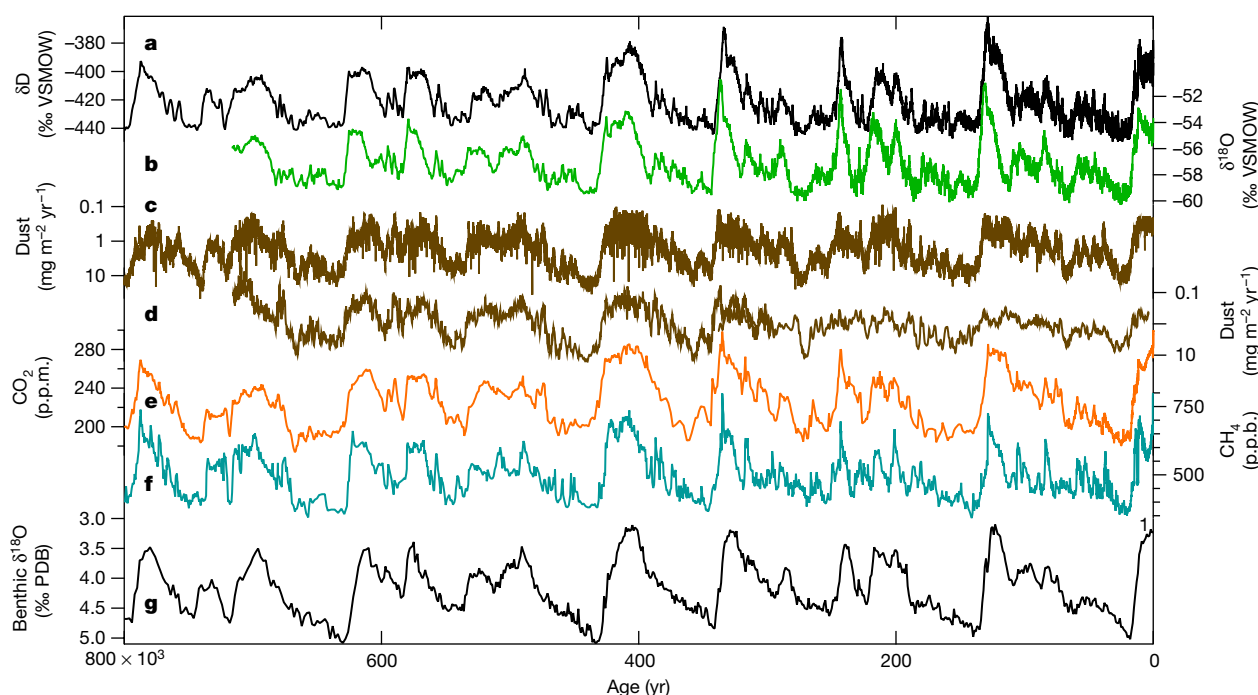
In addition to the palaeotemperature data, the long ice cores provide co-registered records of the flux of dust, sea salt and other atmospheric aerosols, and the concentrations of long-lived atmospheric gases (Fig. 1). Glacial periods are characterized by much higher levels of mineral dust deposition (Fig. 1c and d)<sup>32,33</sup>, commonly attributed to changes in both source strength and atmospheric transport, with the former term dominating<sup>34</sup>. Climate-driven changes in Southern Hemisphere dust source regions (Patagonia and possibly Australia) include more exposure of sources due to aridity or glaciation, and increases in wind strength<sup>32</sup>. Changes in iron delivery to the ocean from dust<sup>32</sup> are probably involved in changes in ocean productivity and atmospheric CO<sub>2</sub> levels<sup>35,36</sup>. There has been considerable interest in developing an (aerosol-based) ice core tracer of past sea ice extent, given the important role of sea ice in the climate system. Initially promising candidates include sea salt sodium and methanesulphonic acid. Further work suggested that quantitative interpretations are difficult because of a variety of confounding effects in the transport and production of both tracers<sup>37</sup>.

The three major greenhouse gases (carbon dioxide (CO<sub>2</sub>), methane (CH<sub>4</sub>) and nitrous oxide (N<sub>2</sub>O)) also show large variations on the same timescale as ice volume and Antarctic climate over the glacial cycles (Fig. 1e and f)<sup>30,38,39</sup>. The largest variations are systematically associated with the glacial terminations, although the imprint of axial tilt and precession is also evident. These greenhouse gas changes act as a positive feedback on the orbitally paced glacial cycles and account for about 40% of the glacial–interglacial change in Earth's radiative balance<sup>10</sup>.

Atmospheric CO<sub>2</sub> concentrations range from roughly 170 to 300 parts per million (p.p.m.) over the glacial–interglacial cycles (Fig. 1e). In addition to their impact on Earth's radiative budget, CO<sub>2</sub> variations reflect global carbon cycling and climate–biosphere interactions. A satisfactory accounting for the full magnitude of the glacial atmospheric CO<sub>2</sub> reduction is still lacking<sup>40,41</sup>, although it is fairly clear that multiple processes operated. Changes in carbon storage on land and changes in CO<sub>2</sub> solubility in the ocean are obvious players, but the sum of these effects results in little net change<sup>40</sup>. Therefore, a higher glacial oceanic carbon inventory must also be involved and both physical and biological processes in the ocean probably contributed.

The Southern Ocean is a key region of interest. The dense, deep water mass originating around Antarctica that fills the abyssal oceans occupied more volume during the last glacial period and was more isolated and poorly ventilated, allowing it to store large quantities of respired carbon<sup>42–44</sup>. Increased efficiency of the biological pump, for example, through iron fertilization<sup>45</sup> or Southern Ocean stratification<sup>46</sup>, could further draw down atmospheric CO<sub>2</sub>. Wind-driven upwelling in the Southern Ocean brings deep waters rich in respired carbon near the surface, allowing carbon exchange with the atmosphere; low nutrient utilization in the Southern Ocean's surface makes this CO<sub>2</sub> 'leak' to the atmosphere more effective<sup>41,46</sup>. Processes that may control the deglacial





**Fig. 1 | Data covering the last 800,000 years from long Antarctic ice core records, and the benthic isotope stack, a proxy for global glacial–interglacial cycles, with upward direction corresponding to warm interglacial conditions. a, The EPICA Dome C  $\delta D$  ( $^2H/^1H$  isotopic ratio of water)<sup>11</sup>. VSMOW, Vienna Standard Mean Ocean Water. b, The Dome**

**Fuji  $\delta^{18}O$  ( $^{18}O/^{16}O$  isotopic ratio of water)<sup>12</sup>. c, d, The EPICA Dome C<sup>81</sup> and Dome Fuji<sup>12</sup> dust records. e, f, The EPICA Dome C/Vostok  $CO_2$ <sup>109</sup> and  $CH_4$ <sup>30</sup> records. g, Benthic oxygen isotope stack<sup>14</sup>. PDB, Pee-Dee belemnite standard.**

release of carbon through the Southern Ocean include the extent of Antarctic summer sea ice<sup>42</sup> or the strength and position of the Southern Hemisphere westerly winds<sup>47</sup>. The atmospheric carbon isotopic composition measured in ice cores is consistent with deglacial ventilation of respired carbon from the deep ocean<sup>48</sup>. The uniformity of glacial minimum  $CO_2$  levels is an interesting challenge, suggesting a consistent negative feedback, the nature of which is not yet clear<sup>49</sup>.

$CH_4$  rose by about 300 parts per billion (p.p.b.) at glacial terminations, with smaller changes during glacial cycles linked to insolation variations driven by orbital tilt and precession<sup>30</sup> (Fig. 1f). On millennial timescales,  $CH_4$  is tightly coupled to Dansgaard–Oeschger events (see section ‘The close view: millennial-scale variability’), although the response appears to be modulated by insolation<sup>50</sup>. A number of factors have been invoked to explain  $CH_4$  variations on these timescales, including changes in emissions from wetlands (the major modern source), release of  $CH_4$  from sea floor hydrates or permafrost, and changes in the atmospheric sink (primarily hydroxyl radical). Isotopic tracers and modelling do not support a dominant role for the last two factors<sup>51–55</sup>. Instead,  $CH_4$  variations are consistent with changes in tropical and mid-latitude hydroclimate, driven by both insolation and shifts in the position of the Inter-Tropical Convergence Zone (ITCZ) during abrupt climate events. Additional boreal contributions may have existed during interglacials when Northern Hemisphere ice sheets retreated<sup>56</sup>. The ice core record contains no indications of very large bursts of  $CH_4$  during interglacial periods, even during those warmer than the Holocene, suggesting that ‘ $CH_4$  time bomb’ scenarios of a runaway positive carbon cycle feedback for the Arctic are unlikely in the near term. A slower, more chronic release of  $CH_4$  from melting permafrost, thermokarst lakes or marine hydrates is expected as a result of global warming<sup>57</sup>.

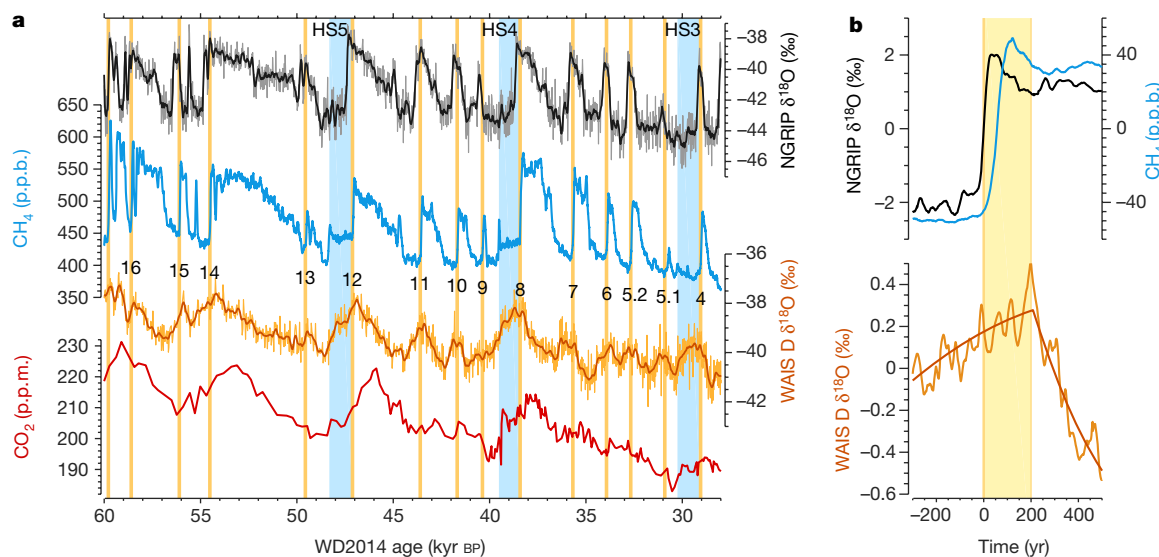
$N_2O$  also varies on glacial–interglacial timescales, with typical glacial and interglacial values of around 210 p.p.b. and 270 p.p.b., respectively<sup>39</sup>. Although the associated radiative forcing is small, the history of this gas is of interest as an integrative tracer of changes in the global nitrogen cycle and because of likely positive feedbacks of global warming on the modern  $N_2O$  budget. The primary natural sources are

microbial nitrification and denitrification in both marine and terrestrial ecosystems. Changes in the ocean source are linked to circulation-driven changes in upper ocean oxygen that affect denitrification<sup>58</sup>. Changes in nitrification on land are linked to temperature and rainfall, primarily in the tropics. Stable isotope data indicate that both the marine and terrestrial sources increased during the last deglaciation, with fast changes<sup>59</sup> in the terrestrial source that paralleled those in  $CH_4$ .

Apart from the changes in the main greenhouse gases, other aspects of the atmospheric evolution revealed by Antarctic ice cores add to our understanding of Earth system change on long timescales; the list below is by no means exhaustive. First, the isotopic composition of atmospheric  $O_2$  shows a strong orbital precession signal, with superimposed millennial-scale variations linked to abrupt climate change. On both timescales this reflects the strength of the global monsoon, which is modulated by both insolation and the position of the ITCZ<sup>13,60</sup>. Second, atmospheric  $O_2$  concentrations have been falling at a rate<sup>61</sup> of 8.4‰ per million years, suggesting that at present  $O_2$  sinks (the oxidation of sedimentary organic carbon and pyrite) exceed  $O_2$  sources (the burial of the same) by about 2%. Third, ice core measurements demonstrate the steady accumulation of radiogenic  $^{40}Ar$  (from  $^{40}K$  decay in the crust) in the atmosphere, allowing an estimate of the contemporary crustal degassing rate<sup>62</sup> and providing a dating technique for old ice. Last, the atmospheric  $Kr/N_2$  ratio is a proxy for global mean ocean temperature<sup>63</sup>, owing to the temperature dependence of solubility. The reduced atmospheric krypton inventory during the Last Glacial Maximum suggests that the mean oceanic temperature was  $2.57 \pm 0.24$  °C lower than at present<sup>64</sup>. The trend of mean ocean warming during the last deglaciation follows Antarctic temperature and atmospheric  $CO_2$  closely, further demonstrating the close link between the high-latitude Southern Hemisphere and the global climate system.

### The close view: millennial-scale variability

Ice cores also preserve the impact of millennial-scale and shorter climate variability, increasingly recognized as important for understanding Earth system feedbacks and the potential for abrupt future change. The now well known, abrupt Dansgaard–Oeschger events



**Fig. 2 | Abrupt climate variability of the last Ice Age. a**, Records of abrupt climate variability. From top to bottom, the traces show Greenland water isotope ratios from the NGRIP core<sup>67</sup>, atmospheric CH<sub>4</sub> from the Antarctic WAIS Divide ice core<sup>85</sup>, Antarctic ice core water isotope ratios from the WAIS Divide core<sup>73</sup> and atmospheric CO<sub>2</sub> from a multi-core compilation<sup>110</sup>. Water isotope ratios are measured relative to VSMOW. Blue bars show the

approximate timing of Heinrich Stadials 5 to 3. Numbers indicate AIM events. **b**, Inter-polar phasing of abrupt climate change. The records from **a** are aligned at the abrupt Northern Hemisphere transitions (yellow vertical lines), and averaged to obtain the shared climatic signal. The Antarctic cooling response of the bipolar seesaw is delayed by around two centuries behind the abrupt Northern Hemisphere events<sup>73</sup>.

observed in Greenlandic ice cores (Fig. 2) and other Northern Hemisphere climate records covering the last ice age<sup>65–67</sup> have well documented counterparts in Antarctica<sup>68–73</sup>, termed Antarctic Isotope Maxima (AIM) events. The precise relative timing of events in Greenland and Antarctica has been established by using well mixed atmospheric gases as stratigraphic markers<sup>68–70,73,74</sup>. Antarctica warmed during Northern Hemisphere cold periods and cooled when Greenland was warm (Fig. 2). This bi-polar linkage is well documented for the last ice age and deglaciation. It very probably operated in prior ice ages but does not have obvious counterparts during interglacials such as the present Holocene. There are regional differences in the expression of the AIM events, which may ultimately provide more information about mechanisms<sup>72</sup>.

The concept of the ‘bi-polar seesaw’ emerged from these observations of asynchronous temperature variations between the hemispheres<sup>27,75</sup>. The basic theory is that perturbations to the northward, cross-equatorial heat transport of the Atlantic Ocean exert opposite temperature effects on both hemispheres, with the Antarctic counterpart damped by a large heat reservoir, commonly assumed to be the Southern Ocean. Climate models in which the strength of the Atlantic Meridional Overturning Circulation (AMOC) is perturbed, usually through freshwater input into the North Atlantic, can reproduce the seesaw pattern<sup>76–78</sup>.

Recently, the WAIS Divide site in West Antarctica provided a climate record with temporal resolution similar to that of the central Greenland cores, although extending only to 68 kyr (the longest stratigraphically ordered Greenland record is NGRIP, at about 123 kyr long). This new ice core combines high accumulation rate, very low gas/ice age difference, and high-precision atmospheric records. These attributes allow for the most precise investigation of the inter-polar phasing of the bipolar seesaw yet<sup>73</sup>. Comparison of Antarctic and Greenlandic events reveals on average an approximately 200-year-long lag of the Antarctic response behind both Greenland abrupt warming and cooling events (Fig. 2). The origin of this delay must lie in the climate coupling between the hemispheres, and indicates a north-to-south propagation of the climate signal dominated by oceanic processes (given that atmospheric propagation would be much faster). Further work on the WAIS Divide ice core<sup>79</sup> explored interhemispheric atmospheric teleconnections by analysing the deuterium excess record, a parameter believed to reflect source moisture conditions and transport pathways to the site.

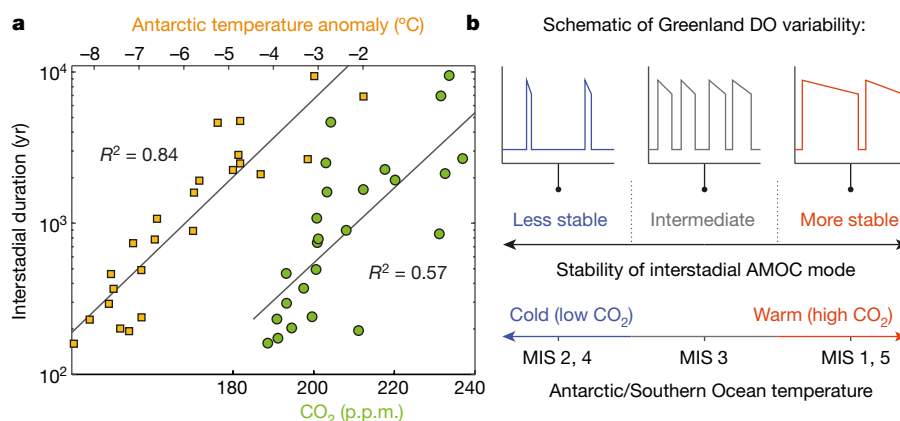
This study showed that a component of the WAIS Divide deuterium excess variations is closely correlated with Greenland climate at zero time lag, implying a fast (atmospheric) link between abrupt warming in the north and shifts in Southern Hemisphere moisture pathways to Antarctica. A similar response had been observed in the Dome C core for termination II<sup>80</sup>. These studies indicate that the two polar regions are coupled via both oceanic and atmospheric teleconnections, each operating on their own timescale.

Tracers of mineral dust deposition in Antarctic ice cores also vary on millennial timescales<sup>6,72,81</sup> with lower dust deposition during warmer periods. Changes in dust transport and generation of dust in source areas are both likely to be involved. Sea salt sodium does not display as clear a link to the millennial-scale climate events, but as discussed above, interpretation of this proxy is complicated. Deposition of both dust and sea salt aerosol appear to be regionally variable within Antarctica.

Atmospheric gas records from Antarctic ice cores are providing more detail about millennial-scale variability in global biogeochemical cycles. Changes in CO<sub>2</sub> tend to follow Antarctic AIM events closely (Fig. 2), particularly for the larger events, as well as for the deglacial warming where the two are essentially synchronous (see Box 2). This close association of Antarctic temperature and CO<sub>2</sub> suggests that Southern Ocean processes are also critical to atmospheric CO<sub>2</sub> variability on these timescales. Changes in export production or ventilation in the Southern Ocean are candidate mechanisms. Both can be linked to AMOC changes, for example, via increased upwelling during Heinrich stadials due to southward migration of the Southern Hemisphere westerly winds<sup>47</sup>, or via more direct impacts of AMOC changes on ocean circulation<sup>82</sup> or the efficiency of the biological pump<sup>58</sup>.

High-resolution CO<sub>2</sub> data reveal that very rapid carbon cycle variations are superimposed on the slower millennial changes during the last ice age and deglaciation<sup>83,84</sup>. There appear to be two types of rapid CO<sub>2</sub> increase. The first is exemplified by increases during two major warming events in the Northern Hemisphere during the deglaciation: the Younger Dryas termination and Bølling–Allerød warming (Box 2). In each case, CO<sub>2</sub> increases of about 10 p.p.m. took place over periods of 100–200 years<sup>84</sup>. Their coincidence with northern warming is unambiguous, given coincident increases in atmospheric CH<sub>4</sub> that mark the two abrupt northern events. The second type is associated with Heinrich stadials, periods of ice-rafted debris discharge into the





**Fig. 3 | Dependence of millennial-scale variability on the background climate state.** **a**, Dansgaard–Oeschger interstadial duration (logarithmic scale) from Greenland ice core records plotted against Antarctic temperature<sup>73</sup> and atmospheric CO<sub>2</sub>, with the coefficient of determination  $R^2$  listed for each case. **b**, Schematic of Greenland Dansgaard–Oeschger variability during various states of the background climate. On the left, during cold (low-CO<sub>2</sub>) periods such as Marine Isotope Stages (MIS) 2 and 4, Dansgaard–Oeschger interstadials are infrequent and of short

duration, suggesting that the interstadial AMOC mode is unstable. In the middle, during intermediate climates such as MIS3, Dansgaard–Oeschger interstadials are frequent and of medium duration, resulting in high event frequency. On the right, during warm (high CO<sub>2</sub>) periods such as MIS 5, Dansgaard–Oeschger interstadials are of long duration, resulting in a lower event frequency, suggesting that the interstadial AMOC mode is very stable. Figure modified from ref. <sup>89</sup> (Wiley).

North Atlantic that occurred during several of the Greenland stadial periods. Clear centennial-scale CO<sub>2</sub> rises are seen during Heinrich Stadial 1 (16.3 kyr BP)<sup>84</sup> (Box 2) and Heinrich Stadial 4 (39.5 kyr BP)<sup>83</sup>. These events are synchronous with short-duration oscillations<sup>85</sup> in atmospheric CH<sub>4</sub>, suggesting that the abrupt shifts in both gases are synchronous with changes in tropical hydrology, for example, due to southward shifts in the position of the ITCZ.

Rapid changes in terrestrial carbon storage are a potential explanation for these abrupt increases in CO<sub>2</sub>. Warming in the Northern Hemisphere at the onset of the Bølling period and end of the Younger Dryas (Box 2) could conceivably alter the carbon balance in terrestrial ecosystems. However, stable isotope data<sup>48</sup> do not support this as a primary mechanism for the CO<sub>2</sub> increases associated with northern warming, and instead suggest that sea surface temperature change is probably at least part of the explanation. The 16.3-kyr CO<sub>2</sub> rise during HS1 is associated with a negative carbon isotopic excursion, implying the release of respired carbon<sup>48</sup>. Drying in the tropics and release of respired carbon associated with the southward migration of the ITCZ could explain this rapid CO<sub>2</sub> change and an analogous CO<sub>2</sub> change during HS4 and possibly HS5 (Fig. 2). This hypothesis is consistent with the suggestion that the small increases in atmospheric CH<sub>4</sub> registered at precisely the same time as the HS1 and HS4 CO<sub>2</sub> increases were caused by a southward shift of ITCZ rainfall<sup>85</sup>.

At least one further aspect of Antarctic ice core data on short timescales deserves mention—the record of volcanism provided by anomalies in non-marine salt sulphate and the presence of volcanic debris (tephra). Ice core records provide highly detailed information about the eruption frequency needed to understand the role of volcanic forcing on climate. Much progress has been made recently in refining these records. For example, identification of synchronous volcanic sulphate deposition in the Antarctic and Greenland indicates tropical eruptions with the potential to affect climate globally<sup>86</sup>. Using the new WAIS Divide record and new data from Greenland, ref. <sup>86</sup> identified five eruptions in the last 2,500 years that were larger than the 1815 Tambora event. Volcanic events also provide critical stratigraphic links that are improving ice core timescales<sup>87,88</sup>, a critical factor for making a unified Antarctic ice core chronology that will allow examination of regional differences in climate.

### The middle ground: orbital–millennial interaction

Commonly, orbital-scale and millennial-scale climate change are treated and discussed in the literature as separate topics. For example, in their seminal paper on the bipolar seesaw, Stocker and Johnson<sup>27</sup>

start by filtering out the orbital signal in order to investigate the abrupt Dansgaard–Oeschger events. Likewise, studies that seek to understand the link between ice volume and insolation (see ref. <sup>19</sup> for example) do not always consider the influence of abrupt events. This approach is of course valid, and much has been learned by studying these timescales in isolation. However, capturing the full dynamics of the climate system requires consideration of the myriad ways in which millennial- and orbital-scale climate change interact, and influence each other.

For example, the duration of Dansgaard–Oeschger interstadial phases scales strongly with the mean climate state<sup>89</sup>. In Fig. 3a, Dansgaard–Oeschger duration is plotted against Antarctic temperature and CO<sub>2</sub>. During relatively warm, high-CO<sub>2</sub> periods such as MIS 5, Dansgaard–Oeschger interstadials tend to be long, suggesting a greater stability of the interstadial (or strong) AMOC mode; conversely, during cold, low-CO<sub>2</sub> periods such as MIS 2 and 4, Dansgaard–Oeschger interstadials are infrequent and of short duration, suggesting a weak interstadial mode that readily collapses (Fig. 3b). Early studies hypothesized a controlling influence of continental ice mass<sup>90,91</sup>, but this is contradicted by global climate model simulations showing a stronger AMOC overturning upon increasing Laurentide ice volume<sup>92</sup>. Viable explanations for this state dependence (Fig. 3) include sea ice dynamics of the North Atlantic<sup>93,94</sup>, the state of the Southern Ocean<sup>89</sup> and CO<sub>2</sub> levels (Fig. 3a)<sup>12</sup>, none of which are mutually exclusive.

Conversely, the bipolar seesaw appears to play an important part in the orbitally paced glacial cycle. As discussed above, data<sup>95</sup> and models<sup>58</sup> suggest that atmospheric CO<sub>2</sub> levels, thought to be the global amplifier of the glacial cycles, are closely linked to (millennial-scale) changes in ocean circulation. Moreover, speleothem records indicate a strongly weakened East Asian monsoon during all of the last seven glacial terminations, interpreted as a southward shift in the ITCZ driven by North Atlantic cooling and AMOC cessation<sup>20</sup>. In this view, glacial terminations are simply the most powerful realizations of the (millennial-scale) Antarctic AIM events<sup>5</sup>. It thus appears that the bipolar seesaw has an important role in the machinery of glacial cycles, and should be considered in answering questions that are commonly placed in the ‘orbital’ realm, such as the aforementioned problems of the 100-kyr cycle and the interhemispheric climate symmetry at the obliquity and precession timescales.

The interdependence of orbital and millennial-scale climate change highlights the need to consider and develop theories that are applicable to both timescales. Transient climate model simulations of the last deglaciation that incorporate both orbital and millennial-scale AMOC forcings have been highly successful in fitting observations<sup>96</sup>,

however, the freshwater fluxes were still prescribed rather than simulated. Similar coupled ocean–atmosphere climate model experiments are needed on all timescales, although the computational cost of such an endeavour is a challenge.

### Lessons for a warming world from Antarctic ice cores

Perhaps the most fundamental message from ice cores is just how profound the anthropogenic impact on our atmospheric composition has been in the context of long-term natural variability. Levels of the primary greenhouse gases CO<sub>2</sub>, CH<sub>4</sub> and N<sub>2</sub>O, all of which are directly affected by human emissions, are at higher levels than at any time<sup>30,38,39</sup> in the last 800 kyr. At the time of writing, concentrations of these three gases are elevated by about 45%, 155% and 22% over pre-industrial values, respectively. Modern changes in CO<sub>2</sub> are also much more rapid than in the ice core record. The fastest pre-industrial increases in CO<sub>2</sub> were about 0.1 p.p.m. per year<sup>84</sup>, compared to consistent recent growth rates close to 2 p.p.m. per year since the early 1990s<sup>97</sup>. For CH<sub>4</sub>, the modern growth rate is also faster than observed through most of the ice core record, although certain natural abrupt transitions match the speed of the current increase<sup>85,97</sup>. Nitrous oxide is not generally sampled closely enough to make such comparisons, but its past growth rate is unlikely to be as rapid as the current rise. The strong correlation of variations in these gases with climate proxies over the last 800 kyr verifies the importance of the greenhouse effect in global climate. As discussed above, the temperature records themselves confirm the theory of polar amplification<sup>10</sup>, an important feature in future warming with implications for Arctic societies and wildlife.

Recent work shows that rapid changes in climate and global biogeochemical cycles during the last ice age and deglaciation affected both the Southern Hemisphere and the tropics<sup>60,73,84,85</sup>, with a strong imprint on tropical systems during Heinrich stadials<sup>60,85</sup>. However, it is not clear what this tells us about what to expect in the future. On the one hand, much of the major millennial variability in the ice core record occurred under different (that is, glacial) boundary conditions, including lower sea level, larger ice sheets and considerably lower CO<sub>2</sub> levels. The lack of those conditions now suggests that in the near future completely analogous events are unlikely. On the other hand, there is a considerable need to understand the strengths of fast feedbacks in the climate system, including the possibility of rapid mass loss at ice sheet margins, the possibility that Greenland meltwater runoff and surface warming could affect the Atlantic overturning circulation, and the potential for fast changes in the global carbon cycle.

The ice core record suggests a more vigorous and stable AMOC during warmer background climate states (Fig. 3); if this palaeo-observation from the last ice age were to apply to future climates also (which has not been rigorously tested), it would imply that the short-term, transient AMOC weakening driven by freshening of the surface North Atlantic may in the long term be offset by an increase in equilibrium AMOC strength in a warmer world<sup>89,98</sup>. Furthermore, changes in the ITCZ and westerly winds appear to be part of the process that transmits millennial signals to Antarctica. It is therefore important to understand their dynamics and impacts better, considering the potential for future changes in regions currently supporting a major fraction of the global human population.

### The future of Antarctic ice core science

The continuous ice core record to 800 kyr is a remarkable achievement. An extension of this record to earlier times, with a goal of reaching back to 1.5 million years BP, is a major new international priority, with ongoing searches for appropriate sites<sup>99</sup>. Several fundamental questions drive this quest. As discussed, ocean sediment records show that before the so-called Mid-Pleistocene Transition (1,200–800 kyr ago), global climate was dominated by a strong 41-kyr period, the cycle associated with the variations in Earth's tilt. After the Mid-Pleistocene Transition the quasi-100-kyr variability dominated. One fundamental question is whether global temperature was warmer at this time, perhaps resulting in smaller and more mobile ice sheets, and shorter ice age cycles<sup>21</sup>. A second

question is whether Antarctic temperature before the Mid-Pleistocene Transition tracked the benthic isotope record, as it does in later times, or varied on some other timescale<sup>100</sup>. A third question concerns whether changes in the long-term mean atmospheric concentrations of greenhouse gases play a part in changing the frequency of glaciation.

The search for a suitable location at which to drill for very old ice involves major investments in radar remote sensing and rapid-access drills<sup>101–103</sup> with which to test probable sites, and the development of better measurement methods. Several international groups are setting their sights on this goal, and it is likely that more than one record will be needed to confirm results. ‘Snapshots’ of older time periods can also be achieved by shallow drilling in ice margin regions<sup>3</sup>, with the potential for finding ice older than 2 million years<sup>104</sup>.

There is also much more to learn about long-term climate and biogeochemical cycles in the existing 800,000-year record through more detailed measurements, including completion of long isotopic records for greenhouse gases and improved (volcanic) synchronization of ice cores both within Antarctica and from Greenland. Understanding the processes that lead to warm interglacials, the nature of abrupt change throughout the record, and the controls on greenhouse gas variability and its links to climate change are some obvious goals. The Holocene history of Antarctica is also critical. Although we have a fairly comprehensive ice core view of the entire Holocene<sup>29,105</sup>, the data needed to put current global warming in the context of Antarctic changes could be improved<sup>106</sup>. Increasing the spatial coverage of high-quality, well dated records would add a great deal to our understanding of this issue.

One of the largest questions about Antarctica for the near future concerns the stability of the West Antarctic Ice Sheet (WAIS). This ice sheet is believed to be vulnerable to collapse as it is grounded below sea level with a retrograde bedrock slope. A critical question that ice core science might answer is whether the ice sheet collapsed during the last interglacial, when temperatures and sea level were higher than today. No cores in the present WAIS penetrate this period, but given the limited number of drilling projects and the high basal melt rate in the area, this cannot be taken as evidence that the WAIS was not there. Modelling suggests that WAIS collapse could be recorded by distinctive patterns in ice core temperature proxies adjacent to the WAIS, related to altered atmospheric circulation<sup>107</sup>. The Mount Moulton blue ice site, at 2,820 m in Marie Byrd Land<sup>108</sup>, provides the only interglacial stable isotope record from West Antarctica; the data are consistent with collapse scenarios<sup>107</sup>. Given uncertainties in models and interpretations of blue ice sites, and the appropriate boundary conditions for collapse scenarios, this result is by no means definitive. Additional coring in and next to the WAIS will be needed to provide better constraints—sites identified for this purpose include Hercules Dome and Skytrain Ice Rise.

Advancing understanding and measurement of ice core proxies will continue to be important. New techniques, for example, deep-ocean temperature proxies from noble gases, ‘clumped isotopes’ in atmospheric gases, mass-independent isotope fractionation in water, sulphate, oxygen and other systems, water isotope diffusion, palaeo-data assimilation techniques, better measurements and more detail in heavy isotope proxies of dust sources, continuous, centimetre-scale water isotope and gas analysis, more detailed isotopic measurements of greenhouse gases, and a host of other methods hold promise for revealing much more about Antarctic and global environmental change.

Finally, although there are ice core records throughout Antarctica (Box 1), the overall spatial coverage is in fact quite thin considering the size and complexity of the continent, and new drilling will continue to be needed to advance scientific goals. International cooperation in ice core drilling and prioritizing science is strong. The technical expertise required to accomplish challenging field programmes is available in many nations, placing ice core science in an excellent position to improve our understanding of the history of Antarctica and its links to the larger Earth system.

Received: 10 November 2017; Accepted: 19 March 2018;  
Published online 13 June 2018.



1. Galeotti, S. et al. Antarctic Ice Sheet variability across the Eocene-Oligocene boundary climate transition. *Science* **352**, 76–80 (2016).
2. Flower, B. P. & Kennett, J. P. Relations between Monterey Formation deposition and middle Miocene global cooling: Naples Beach section. *Calif. Geol.* **21**, 877–880 (1993).
3. Higgins, J. A. et al. Atmospheric composition 1 million years ago from blue ice in the Allan Hills, Antarctica. *Proc. Natl Acad. Sci. USA* **112**, 6887–6891 (2015). **This study reports the first greenhouse gas data from ice older than 800,000 years.**
4. Jouzel, J. et al. Validity of the temperature reconstruction from water isotopes in ice cores. *J. Geophys. Res. Oceans* **102**, 26471–26487 (1997).
5. Wolff, E., Fischer, H. & Röthlisberger, R. Glacial terminations as southern warmings without northern control. *Nat. Geosci.* **2**, 206–209 (2009).
6. Schüpbach, S. et al. High-resolution mineral dust and sea ice proxy records from the Talos Dome ice core. *Clim. Past* **9**, 2789–2807 (2013).
7. McConnell, J. R. et al. Antarctic-wide array of high-resolution ice core records reveals pervasive lead pollution began in 1889 and persists today. *Sci. Rep.* **4**, 5848 (2014).
8. Brook, E. J., Kurz, M. D. & Curtice, J. Flux and size fractionation of  $^3\text{He}$  in interplanetary dust from Antarctic ice core samples. *Earth Planet. Sci. Lett.* **286**, 565–569 (2009).
9. Van Ommen, T. D., Morgan, V. & Curran, M. A. Deglacial and Holocene changes in accumulation at Law Dome, East Antarctica. *Ann. Glaciol.* **39**, 359–365 (2004).
10. Cuffey, K. M. et al. Deglacial temperature history of West Antarctica. *Proc. Natl Acad. Sci. USA* **113**, 14249–14254 (2016). **This work provides the first accurate borehole-based temperature reconstruction from Antarctica, indicating a glacial–interglacial temperature change of  $11.3 \pm 1.8^\circ\text{C}$ .**
11. Jouzel, J. et al. Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* **317**, 793–796 (2007). **This paper reports the full 800,000-year temperature reconstruction from the EPICA Dome C ice core, the longest such record.**
12. Dome Fuji Project Members. State dependence of climatic instability over the past 720,000 years from Antarctic ice cores and climate modeling. *Sci. Adv.* **3**, e1600446 (2017).
13. Petit, J. R. et al. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
14. Lisiecki, L. E. & Raymo, M. E. A Pliocene-Pleistocene stack of 57 globally distributed benthic  $\delta^{18}\text{O}$  records. *Paleoceanography* **20**, 1–17 (2005).
15. Shakun, J. D. et al. Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nat. Geosci.* **4**, 49–55 (2012).
16. Hays, J. D., Imbrie, J. & Shackleton, N. J. Variations in the Earth's orbit: pacemaker of the ice ages. *Science* **194**, 1121–1132 (1976).
17. Imbrie, J. et al. On the structure and origin of major glaciation cycles. *Paleoceanogr. Paleoclimatol.* **8**, 699–735 (1993).
18. Raymo, M. The timing of major climate terminations. *Paleoceanogr. Paleoclimatol.* **12**, 577–585 (1997).
19. Paillard, D. The timing of Pleistocene glaciations from a simple multiple-state climate model. *Nature* **391**, 378–381 (1998).
20. Cheng, H. et al. The Asian monsoon over the past 640,000 years and ice age terminations. *Nature* **534**, 640–646 (2016).
21. Tzedakis, P., Crucifix, M., Mitsui, T. & Wolff, E. W. A simple rule to determine which insolation cycles lead to interglacials. *Nature* **542**, 427–432 (2017).
22. Bintanja, R. & Van de Wal, R. North American ice-sheet dynamics and the onset of 100,000-year glacial cycles. *Nature* **454**, 869–872 (2008).
23. Abe-Ouchi, A. et al. Insolation-driven 100,000-year glacial cycles and hysteresis of ice-sheet volume. *Nature* **500**, 190–193 (2013).
24. Ganopolski, A. & Brovkin, V. Simulation of climate, ice sheets and  $\text{CO}_2$  evolution during the last four glacial cycles with an Earth system model of intermediate complexity. *Clim. Past* **13**, 1695–1716 (2017).
25. Broecker, W. S. & Denton, G. H. The role of ocean-atmosphere reorganizations in glacial cycles. *Quat. Sci. Rev.* **9**, 305–341 (1990).
26. Kawamura, K. et al. Northern Hemisphere forcing of climatic cycles in Antarctica over the past 360,000 years. *Nature* **448**, 912–916 (2007). **This study links variations in the  $\text{O}_2/\text{N}_2$  ratio of trapped air in the Dome Fuji ice core to local summer insolation, thereby dating the core and showing that glacial terminations in Antarctica closely followed Northern Hemisphere summer insolation.**
27. Stocker, T. F. & Johnsen, S. J. A minimum thermodynamic model for the bipolar seesaw. *Paleoceanogr. Paleoclimatol.* **18**, <https://doi.org/10.1029/2003PA000920> (2003).
28. Huybers, P. & Denton, G. Antarctic temperature at orbital timescales controlled by local summer duration. *Nat. Geosci.* **1**, 787–792 (2008).
29. WAIS Divide Project Members. Onset of deglacial warming in West Antarctica driven by local orbital forcing. *Nature* **500**, 440–444 (2013).
30. Loulergue, L. et al. Orbital and millennial-scale features of atmospheric  $\text{CH}_4$  over the past 800,000 years. *Nature* **453**, 383–386 (2008). **This paper reports the full 800,000-year atmospheric methane record from the EPICA Dome C ice core, showing variations on orbital and millennial timescales.**
31. Yin, Q. Insolation-induced mid-Brunhes transition in Southern Ocean ventilation and deep-ocean temperature. *Nature* **494**, 222–225 (2013).
32. Wolff, E. W. et al. Southern Ocean sea-ice extent, productivity and iron flux over the past eight glacial cycles. *Nature* **440**, 491–496 (2006). **Chemical measurements from the EPICA Dome C ice core indicate that both the flux of iron (from wind-blown dust) and sea ice extent increased during glacial periods over the past 740,000 years.**
33. Lambert, F. et al. Dust–climate couplings over the past 800,000 years from the EPICA Dome C ice core. *Nature* **452**, 616–619 (2008).
34. Fischer, H., Siggaard-Andersen, M. L., Ruth, U., Röthlisberger, R. & Wolff, E. Glacial/interglacial changes in mineral dust and sea-salt records in polar ice cores: sources, transport, and deposition. *Rev. Geophys.* **45**, 1–26 (2007).
35. Martínez-García, A. et al. Links between iron supply, marine productivity, sea surface temperature, and  $\text{CO}_2$  over the last 1.1 Ma. *Paleoceanogr. Paleoclimatol.* **24**, PA1207 (2009).
36. Jaccard, S. et al. Two modes of change in Southern Ocean productivity over the past million years. *Science* **339**, 1419–1423 (2013).
37. Abram, N. J., Wolff, E. W. & Curran, M. A. A review of sea ice proxy information from polar ice cores. *Quat. Sci. Rev.* **79**, 168–183 (2013).
38. Lüthi, D. et al. High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* **453**, 379–382 (2008). **This paper completed the 800,000-year-long EPICA Dome C  $\text{CO}_2$  record shown in Fig. 1, demonstrating the variation of  $\text{CO}_2$  maxima during interglacial times.**
39. Schilt, A. et al. Glacial-interglacial and millennial-scale variations in the atmospheric nitrous oxide concentration during the last 800,000 years. *Quat. Sci. Rev.* **29**, 182–192 (2010).
40. Sigman, D. M. & Boyle, E. A. Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* **407**, 859–869 (2000).
41. Sigman, D. M., Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in atmospheric  $\text{CO}_2$  concentration. *Nature* **466**, 47–55 (2010).
42. Stephens, B. B. & Keeling, R. F. The influence of Antarctic sea ice on glacial-interglacial  $\text{CO}_2$  variations. *Nature* **404**, 171–174 (2000).
43. Skinner, L., Fallon, S., Waelbroeck, C., Michel, E. & Barker, S. Ventilation of the deep Southern Ocean and deglacial  $\text{CO}_2$  rise. *Science* **328**, 1147–1151 (2010).
44. Ferrari, R. et al. Antarctic sea ice control on ocean circulation in present and glacial climates. *Proc. Natl Acad. Sci. USA* **111**, 8753–8758 (2014).
45. Martin, J. H. Glacial-interglacial  $\text{CO}_2$  change: the iron hypothesis. *Paleoceanogr. Paleoclimatol.* **5**, 1–13 (1990).
46. Franois, R. et al. Contribution of Southern Ocean surface-water stratification to low atmospheric  $\text{CO}_2$  concentrations during the last glacial period. *Nature* **389**, 929–935 (1997).
47. Anderson, R. et al. Wind-driven upwelling in the Southern Ocean and the deglacial rise in atmospheric  $\text{CO}_2$ . *Science* **323**, 1443–1448 (2009).
48. Bauska, T. K. et al. Carbon isotopes characterize rapid changes in atmospheric carbon dioxide during the last deglaciation. *Proc. Natl Acad. Sci. USA* **113**, 3465–3470 (2016).
49. Galbraith, E. & Eggleston, S. A lower limit to atmospheric  $\text{CO}_2$  concentrations over the past 800,000 years. *Nat. Geosci.* **10**, 295–298 (2017).
50. Brook, E. J., Sowers, T. & Orchard, J. Rapid variations in atmospheric methane concentration during the past 110,000 years. *Science* **273**, 1087–1091 (1996).
51. Petrenko, V. V. et al. Minimal geological methane emissions during the Younger Dryas–Preboreal abrupt warming event. *Nature* **548**, 443–446 (2017).
52. Sowers, T. Late quaternary atmospheric  $\text{CH}_4$  isotope record suggests marine clathrates are stable. *Science* **311**, 838–840 (2006).
53. Bock, M. et al. Hydrogen isotopes preclude marine hydrate  $\text{CH}_4$  emissions at the onset of Dansgaard-Oeschger events. *Science* **328**, 1686–1689 (2010).
54. Levine, J. et al. Reconciling the changes in atmospheric methane sources and sinks between the Last Glacial Maximum and the pre-industrial era. *Geophys. Res. Lett.* **38**, L23804 (2011).
55. Murray, L. T. et al. Factors controlling variability in the oxidative capacity of the troposphere since the Last Glacial Maximum. *Atmos. Chem. Phys.* **14**, 3589–3622 (2014).
56. Fischer, H. et al. Changing boreal methane sources and constant biomass burning during the last termination. *Nature* **452**, 864–867 (2008).
57. Brook, E., Archer, D., Dlugokencky, E., Frolking, S. & Lawrence, D. In *Abundant Climate Change. A report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research* (ed. McGeehin, J. P.) Ch. 5, 163–201 (US Geological Survey, Reston, 2008).
58. Schmittner, A. & Galbraith, E. D. Glacial greenhouse-gas fluctuations controlled by ocean circulation changes. *Nature* **456**, 373–376 (2008).
59. Schilt, A. et al. Isotopic constraints on marine and terrestrial  $\text{N}_2\text{O}$  emissions during the last deglaciation. *Nature* **516**, 234–237 (2014).
60. Severinghaus, J. P., Beaudette, R., Headly, M. A., Taylor, K. & Brook, E. J. Oxygen-18 of  $\text{O}_2$  records the impact of abrupt climate change on the terrestrial biosphere. *Science* **324**, 1431–1434 (2009).
61. Stolper, D., Bender, M., Dreyfus, G., Yan, Y. & Higgins, J. A Pleistocene ice core record of atmospheric  $\text{O}_2$  concentrations. *Science* **353**, 1427–1430 (2016).
62. Bender, M. L., Barnett, B., Dreyfus, G., Jouzel, J. & Porcelli, D. The contemporary degassing rate of  $^{40}\text{Ar}$  from the solid Earth. *Proc. Natl Acad. Sci. USA* **105**, 8232–8237 (2008). **Precise measurements of argon isotope ratios in trapped air are used in this study to develop a chronometer for old ice based on the accumulation of  $^{40}\text{Ar}$  in the atmosphere from  $^{40}\text{K}$  decay in the crust.**
63. Headly, M. A. & Severinghaus, J. P. A method to measure  $\text{Kr}/\text{N}_2$  ratios in air bubbles trapped in ice cores and its application in reconstructing past mean ocean temperature. *J. Geophys. Res.* **112**, D19105 (2007).
64. Bereiter, B., Shackleton, S., Baggenstos, D., Kawamura, K. & Severinghaus, J. Mean global ocean temperatures during the last glacial transition. *Nature* **553**, 39–44 (2018). **Measurements of Kr, Xe, Ar, and N are used in this study to make the first precise estimates of changes in global deep-ocean temperature across the last glacial-interglacial transition, showing that these are mostly synchronous with changes in Antarctic air temperature and atmospheric  $\text{CO}_2$ .**

65. Grootes, P., Stuiver, M., White, J., Johnsen, S. & Jouzel, J. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature* **366**, 552–554 (1993).
66. Dansgaard, W. et al. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**, 218–220 (1993).  
**This paper reports the first detailed record of abrupt changes in temperature in Greenland.**
67. Andersen, K. K. et al. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431**, 147–151 (2004).
68. Blunier, T. & Brook, E. J. Timing of millennial-scale climate change in Antarctica and Greenland during the last glacial period. *Science* **291**, 109–112 (2001).  
**This paper used methane variations to make a common timescale for ice cores in Greenland and Antarctica, and showed the bi-polar seesaw pattern for the larger climate variations during the last ice age.**
69. Brook, E. J. et al. Timing of millennial-scale climate change at Siple Dome, West Antarctica, during the last glacial period. *Quat. Sci. Rev.* **24**, 1333–1343 (2005).
70. EPICA Community Members. One-to-one coupling of glacial climate variability in Greenland and Antarctica. *Nature* **444**, 195–198 (2006).  
**This paper demonstrated that all abrupt climate events in the Greenland record have counterparts in Antarctica.**
71. Stenni, B. et al. Expression of the bipolar see-saw in Antarctic climate records during the last deglaciation. *Nat. Geosci.* **4**, 46–49 (2011).
72. Landais, A. et al. A review of the bipolar see-saw from synchronized and high resolution ice core water stable isotope records from Greenland and East Antarctica. *Quat. Sci. Rev.* **114**, 18–32 (2015).
73. WAIS Divide Project Members. Precise inter-core phasing of abrupt climate change during the last ice age. *Nature* **520**, 661–665 (2015).  
**Using very precise chronological constraints, this paper demonstrated that millennial-scale warming and cooling in Antarctica lagged counterpart events in Greenland by about 200 years on average.**
74. Bender, M. et al. Climate correlations between Greenland and Antarctica during the past 100,000 years. *Nature* **372**, 663–666 (1994).
75. Broecker, W. S. Paleocene circulation during the last deglaciation: a bipolar seesaw? *Paleoceanogr. Paleoclimatol.* **13**, 119–121 (1998).
76. Vellinga, M. & Wood, R. A. Global climatic impacts of a collapse of the Atlantic thermohaline circulation. *Clim. Change* **54**, 251–267 (2002).
77. Schmittner, A., Saenko, O. & Weaver, A. Coupling of the hemispheres in observations and simulations of glacial climate change. *Quat. Sci. Rev.* **22**, 659–671 (2003).
78. Stouffer, R. J. et al. Investigating the causes of the response of the thermohaline circulation to past and future climate changes. *J. Clim.* **19**, 1365–1387 (2006).
79. Markle, B. R. et al. Global atmospheric teleconnections during Dansgaard-Oeschger events. *Nat. Geosci.* **10**, 36–40 (2017).
80. Masson-Delmotte, V. et al. Abrupt change of Antarctic moisture origin at the end of Termination II. *Proc. Natl Acad. Sci. USA* **107**, 12091–12094 (2010).
81. Lambert, F., Bigler, M., Steffensen, J. P., Hutterli, M. & Fischer, H. Centennial mineral dust variability in high-resolution ice core data from Dome C, Antarctica. *Clim. Past* **8**, 609–623 (2012).
82. Menviel, L., England, M. H., Meissner, K., Mouchet, A. & Yu, J. Atlantic-Pacific seesaw and its role in outgassing CO<sub>2</sub> during Heinrich events. *Paleoceanogr. Paleoclimatol.* **29**, 58–70 (2014).
83. Ahn, J., Brook, E. J., Schmittner, A. & Kreutz, K. Abrupt change in atmospheric CO<sub>2</sub> during the last ice age. *Geophys. Res. Lett.* **39**, GL053018 (2012).
84. Marcott, S. A. et al. Centennial-scale changes in the global carbon cycle during the last deglaciation. *Nature* **514**, 616–619 (2014).  
**The most detailed CO<sub>2</sub> record for the deglaciation to date is reported from the WAIS Divide ice core in this paper, showing the tight coupling of CO<sub>2</sub> and Antarctic climate (Box 2).**
85. Rhodes, R. H. et al. Enhanced tropical methane production in response to iceberg discharge in the North Atlantic. *Science* **348**, 1016–1019 (2015).
86. Sigl, M. et al. Timing and climate forcing of volcanic eruptions for the past 2,500 years. *Nature* **523**, 543–549 (2015).
87. Veres, D. et al. The Antarctic ice core chronology (AICC2012): an optimized multi-parameter and multi-site dating approach for the last 120 thousand years. *Clim. Past* **9**, 1733–1748 (2013).
88. Bazin, L. et al. An optimized multi-proxy, multi-site Antarctic ice and gas orbital chronology (AICC2012): 120–800 ka. *Clim. Past Discuss.* **8**, 5963–6009 (2012).
89. Buizert, C. & Schmittner, A. Southern Ocean control of glacial AMOC stability and Dansgaard-Oeschger interstadial duration. *Paleoceanogr. Paleoclimatol.* **30**, 1595–1612 (2015).
90. McManus, J. F., Oppo, D. W. & Cullen, J. L. A. 0.5-million-year record of millennial-scale climate variability in the North Atlantic. *Science* **283**, 971–975 (1999).
91. Schulz, M., Berger, W. H., Sarnthein, M. & Grootes, P. M. Amplitude variations of 1470-year climate oscillations during the last 100,000 years linked to fluctuations of continental ice mass. *Geophys. Res. Lett.* **26**, 3385–3388 (1999).
92. Muglia, J. & Schmittner, A. Glacial Atlantic overturning increased by wind stress in climate models. *Geophys. Res. Lett.* **42**, 9862–9868 (2015).
93. Oka, A., Hasumi, H. & Abe-Ouchi, A. The thermal threshold of the Atlantic meridional overturning circulation and its control by wind stress forcing during glacial climate. *Geophys. Res. Lett.* **39**, GL051421 (2012).
94. Wang, Z. & Mysak, L. A. Glacial abrupt climate changes and Dansgaard-Oeschger oscillations in a coupled climate model. *Paleoceanogr. Paleoclimatol.* **21**, PA001238 (2006).
95. Ahn, J. & Brook, E. J. Atmospheric CO<sub>2</sub> and climate on millennial time scales during the last glacial period. *Science* **322**, 83–85 (2008).
96. Liu, Z. et al. Transient simulation of last deglaciation with a new mechanism for Bølling-Allerød warming. *Science* **325**, 310–314 (2009).
97. Dlugokencky, E. *Trends in Atmospheric Methane*. [http://www.esrl.noaa.gov/gmd/ccgg/trends\\_ch4/](http://www.esrl.noaa.gov/gmd/ccgg/trends_ch4/) (Earth System Research Laboratory, 2018).
98. Toggweiler, J. & Russell, J. Ocean circulation in a warming climate. *Nature* **451**, 286–288 (2008).
99. Fischer, H. et al. Where to find 1.5 million yr old ice for the IPICS “Oldest-Ice” ice core. *Clim. Past* **9**, 2489–2505 (2013).
100. Raymo, M., Lisiecki, L. & Nisancioglu, K. H. Plio-Pleistocene ice volume, Antarctic climate, and the global  $\delta^{18}\text{O}$  record. *Science* **313**, 492–495 (2006).
101. Schwander, J., Marending, S., Stocker, T. & Fischer, H. RADIX: a minimal-resources rapid-access drilling system. *Ann. Glaciol.* **55**, 34–38 (2014).
102. Alemany, O. et al. The SUBGLACIOR drilling probe: concept and design. *Ann. Glaciol.* **55**, 233–242 (2014).
103. Goodge, J. W. & Severinghaus, J. P. Rapid Access Ice Drill: a new tool for exploration of the deep Antarctic ice sheets and subglacial geology. *J. Glaciol.* **62**, 1049–1064 (2016).
104. Yan, Y. N. J. et al. 2.7-Million-Year-Old Ice from Allan Hills Blue Ice Areas, East Antarctica Reveals Climate Snapshots Since Early Pleistocene. *Goldschmidt Conf. (Paris, France)* 4359, <https://goldschmidtabstracts.info/2017/4359.pdf> (European Association of Geochemistry and the Geochemical Society, 2007).
105. Masson-Delmotte, V. et al. A comparison of the present and last interglacial periods in six Antarctic ice cores. *Clim. Past* **7**, 397–423 (2011).
106. Pages 2k Consortium. Continental-scale temperature variability during the past two millennia. *Nat. Geosci.* **6**, 339–350 (2013).
107. Steig, E. J. et al. Influence of West Antarctic Ice Sheet collapse on Antarctic surface climate. *Geophys. Res. Lett.* **42**, 4862–4868 (2015).
108. Korotkiy, E. V. et al. The last interglacial as represented in the glaciochemical record from Mount Moulton Blue Ice Area, West Antarctica. *Quat. Sci. Rev.* **30**, 1940–1947 (2011).
109. Bereiter, B. et al. Revision of the EPICA Dome C CO<sub>2</sub> record from 800 to 600 kyr before present. *Geophys. Res. Lett.* **42**, 542–549 (2015).
110. Bereiter, B. et al. Mode change of millennial CO<sub>2</sub> variability during the last glacial cycle associated with a bipolar marine carbon seesaw. *Proc. Natl Acad. Sci. USA* **109**, 9755–9760 (2012).
111. Gow, A. J., Ueda, H. T. & Garfield, D. E. Antarctic ice sheet: preliminary results of first core hole to bedrock. *Science* **161**, 1011–1013 (1968).
112. EPICA Community Members. Eight glacial cycles from an Antarctic ice core. *Nature* **429**, 623–628 (2004).
113. Slawny, K. R. et al. Production drilling at WAIS Divide. *Ann. Glaciol.* **55**, 147–155 (2014).
114. Neftel, A., Oeschger, H., Staffelbach, T. & Stauffer, B. CO<sub>2</sub> record in the Byrd ice core 50,000–5,000 years BP. *Nature* **331**, 609–611 (1988).
115. Barnola, J. M., Pimienta, P., Raynaud, D. & Korotkevich, Y. S. CO<sub>2</sub>-climate relationship as deduced from the Vostok Ice Core—a reexamination based on new measurements and on a reevaluation of the air dating. *Tellus B* **43**, 83–90 (1991).
116. Fischer, H., Wahlen, M., Smith, J., Mastropianni, D. & Deck, B. Ice core records of atmospheric CO<sub>2</sub> around the last three glacial terminations. *Science* **283**, 1712–1714 (1999).
117. Monnin, E. et al. Atmospheric CO<sub>2</sub> concentrations over the last glacial termination. *Science* **291**, 112–114 (2001).
118. Pedro, J. B., Rasmussen, S. O. & van Ommen, T. D. Tightened constraints on the time-lag between Antarctic temperature and CO<sub>2</sub> during the last deglaciation. *Clim. Past* **8**, 1213–1221 (2012).
119. Parrenin, F. et al. Synchronous change of atmospheric CO<sub>2</sub> and Antarctic temperature during the last deglacial warming. *Science* **339**, 1060–1063 (2013).

**Acknowledgements** We thank J. Pedro for comments that improved the manuscript. The US National Science Foundation and US Antarctic Program have provided support for our research and acquisition of Antarctic ice cores that we have studied; we thank them, as well as numerous international agencies and colleagues who have contributed to ice core science.

**Reviewer information** Nature thanks J. Pedro and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** The authors contributed equally to this work.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to E.J.B.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# The global influence of localized dynamics in the Southern Ocean

Stephen R. Rintoul<sup>1\*</sup>

**The circulation of the Southern Ocean connects ocean basins, links the deep and shallow layers of the ocean, and has a strong influence on global ocean circulation, climate, biogeochemical cycles and the Antarctic Ice Sheet. Processes that act on local and regional scales, which are often mediated by the interaction of the flow with topography, are fundamental in shaping the large-scale, three-dimensional circulation of the Southern Ocean. Recent advances provide insight into the response of the Southern Ocean to future change and the implications for climate, the carbon cycle and sea-level rise.**

The ocean circles the globe in the latitude band of Drake Passage (56° S–58° S), unblocked by continents. As a consequence of this unique geometry, the Southern Ocean influences the ocean circulation and climate on global scales. The multiple jets of the Antarctic Circumpolar Current (ACC), the largest ocean current, flow from west to east through this channel and connect the ocean basins (Fig. 1a). The unblocked channel thus enhances interbasin exchange, but inhibits north–south exchange because there can be no net meridional geostrophic flow in the absence of zonal pressure gradients supported by land barriers or topography. Surfaces of constant density rise steeply to the south, in geostrophic balance with the strong eastward flow of the circumpolar current. The steeply sloping isopycnals (see Box 1 for a glossary of terms used) provide a reservoir of potential energy that is extracted by baroclinic instability to drive the vigorous eddy field that is evident in Fig. 1a. Eddies, in turn, transport fluid and tracers across the ACC. In particular, deep water spreads polewards and rises along the sloping isopycnals, reaching the sea surface near Antarctica<sup>1,2</sup> (Fig. 1b). Vigorous interactions between the atmosphere, ocean and cryosphere transform upwelled deep waters into either dense Antarctic Bottom Water (AABW) or lighter intermediate waters (Subantarctic Mode Water (SAMW) or Antarctic Intermediate Water (AAIW)). The result of these water-mass transformations is an overturning circulation that consists of two cells: an upper cell in which dense deep water is converted to lighter waters, and a lower cell in which deep water is converted to denser bottom water<sup>1–4</sup> (Fig. 1b). The poleward flow of deep water is balanced by equatorward flow of intermediate and bottom waters formed in the Southern Ocean. By connecting the ocean basins and the deep and shallow layers of the ocean, the Southern Ocean circulation allows a global-scale ocean overturning in which conversion of deep water to intermediate water in the Southern Ocean largely compensates the sinking of deep water in the North Atlantic<sup>4,5</sup>. The global overturning circulation, in turn, largely sets the capacity of the ocean to store and transport heat and carbon dioxide and thereby influence climate<sup>6–8</sup>.

Several characteristics set the circulation of the Southern Ocean apart from ocean currents in other regions. Weak stratification and strong eastward flow driven by powerful winds over the Southern Ocean combine to establish momentum and vorticity balances that differ from those at lower latitudes. In particular, the adjustment process that limits the depth of the wind-driven circulation at lower latitudes does not operate in the ACC<sup>9</sup>. In the subtropics, changes in wind generate planetary waves that propagate west and gradually establish a

new equilibrium circulation of the gyres found in the upper kilometre or so of the water column<sup>10</sup>. In the ACC, the eastward flow is much faster than the westward propagation of the waves. As a consequence, the current extends to great depth and the flow is strongly influenced by topography.

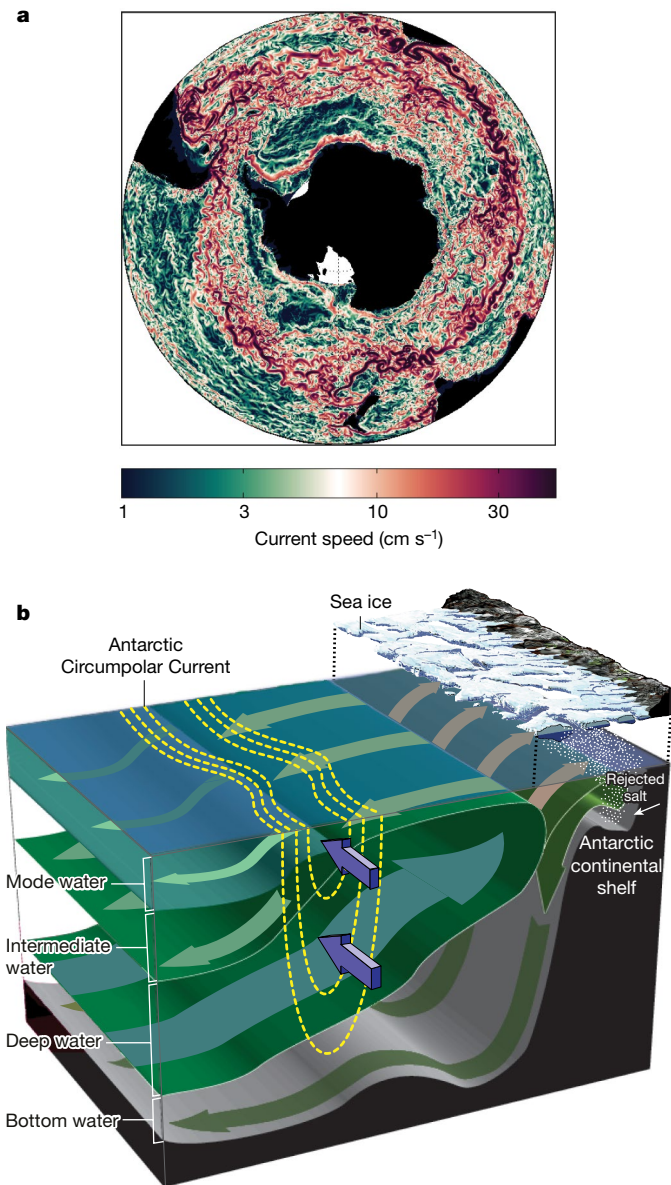
Substantial progress in understanding the dynamics of the Southern Ocean and its role in the climate system has been made by adopting the zonally averaged perspective shown schematically in Fig. 1b. Here I describe new insights gained from observations, theory and models that highlight the limitations of the zonal-mean view and the importance of local and regional dynamics. ('Local' in this context refers to processes that are initiated in a particular area, over spatial scales of tens to hundreds of kilometres, but whose influence on the flow may extend hundreds or thousands of kilometres downstream.) Recent studies have also elucidated the pathways responsible for sequestering heat and carbon and their unanticipated sensitivity to fluctuations in climate forcing. The processes that drive the overturning circulation are better understood, including appreciation of the contribution of fresh-water transport by sea-ice formation and melt. The extent to which the fate of the Antarctic Ice Sheet is linked to changes in the surrounding ocean has come into sharper focus in the past decade. As observational records have increased in length and coverage, the nature and drivers of variability and change in the Southern Ocean have become better understood. Taken together, these recent developments provide the foundation for a new conceptual model of the Southern Ocean, in which the large-scale circulation that is of such importance to global climate emerges from dynamics that play out on local and regional scales, catalysed by topography.

## Zonal-average dynamics of the Southern Ocean

As illustrated in Fig. 1, the circulation of the Southern Ocean consists of two main elements: the strong eastward flow of the ACC and a weaker overturning circulation that carries water towards or away from the Antarctic continent. By the early 2000s, sufficient progress had been made to allow articulation of a dynamical recipe for these two dominant aspects of the Southern Ocean circulation and their interaction, as summarized in recent reviews<sup>11–13</sup>. The essential ingredients include a circumpolar channel with realistic bathymetry, driven at the surface by wind and buoyancy forcing. Strong westerly winds drive the northward transport of surface waters in the Ekman layer, with convergence (downwelling) north of the wind-stress maximum and

<sup>1</sup>CSIRO Oceans and Atmosphere, Antarctic Climate and Ecosystems Cooperative Research Centre, Centre for Southern Hemisphere Ocean Research, Hobart, Tasmania, Australia.

\*e-mail: [steve.rintoul@csiro.au](mailto:steve.rintoul@csiro.au)



**Fig. 1 | The Southern Ocean circulation.** **a**, Five-day-mean current speed at a depth of 112 m from a high-resolution ( $1/12^\circ$ ) numerical simulation of the ACC, illustrating the filamented, eddy-rich structure of the current. Image adapted from ref. <sup>110</sup> (movie S1 in the supporting information), American Geophysical Union. **b**, A highly schematic illustration of the Southern Ocean overturning circulation. Deep water spreads polewards and upwards across the ACC. Water that upwells close to Antarctica is converted to denser bottom water by cooling and brine released during sea-ice formation (indicated by white dots in the upper right of the diagram). Water that upwells further north is converted to lighter mode and intermediate waters that sink to the north of the current and ventilate the intermediate depths of the ocean. The schematic omits many important aspects of the Southern Ocean circulation, including wind and buoyancy forcing, eddies and mixing processes. Image adapted from ref. <sup>111</sup> (<https://doi.org/10.26749/rstpp.133.3.41>, figure 6), Royal Society of Tasmania.

divergence (upwelling) south of it<sup>1,2</sup>. The winds therefore cause isopycnals to slope upwards to the south, establishing an eastward geostrophic current. Once the isopycnals are sufficiently steep, they become baroclinically unstable and form eddies<sup>12</sup>. The vigorous eddy field plays a central part in Southern Ocean dynamics. Eddies transfer momentum vertically from the sea surface to the sea floor<sup>11,12</sup>, where bottom-form stress balances the wind stress<sup>14</sup>. Downward flux of momentum by eddies is associated with poleward transport of heat, and eddies make the dominant contribution to the poleward heat flux that is needed to

balance the heat lost to the atmosphere at high latitudes<sup>15</sup>. Buoyancy forcing can also influence ACC transport by altering the stratification and cross-stream density gradient<sup>12,16</sup>.

Although early measurements of temperature and salinity were sufficient to reveal the existence of an overturning circulation that carries saline deep water towards Antarctica and a return flow of fresher waters near the sea floor and in the upper ocean<sup>1,2</sup> (Fig. 1b), the dynamics of the overturning and its connection to the ACC remained obscure for many decades. Progress in the late twentieth century revealed that the overturning circulation is in fact intimately linked to the dynamics of the ACC and its eddy field. Because eddies carry mass polewards, allowing meridional transport across the unbounded channel of the Southern Ocean at depths above the shallowest topography, the eddy field associated with the unstable flow of the ACC is connected directly to the overturning circulation. In a stably stratified ocean, a zonal-mean overturning circulation can exist only if buoyancy is added or removed to convert water from one density class to another (for example, conversion of dense deep water to lighter mode and intermediate water in the upper cell requires an input of buoyancy, whereas conversion of deep water to denser bottom water in the lower cell requires removal of buoyancy). As a consequence, the strength of the overturning is related directly to the buoyancy forcing at the sea surface<sup>3,17</sup>.

A key question is how the circulation of the Southern Ocean, both the ACC and the overturning, responds to changes in forcing by the atmosphere (that is, changes in wind stress at the sea surface or in air-sea exchange of heat and moisture). Two concepts have dominated recent discussion of the response of the Southern Ocean to changes in forcing: ‘eddy saturation’ and ‘eddy compensation’. In the eddy-saturation limit, stronger wind forcing results in a more vigorous eddy field, with little change in ACC transport<sup>18</sup>. Observations of little change in the isopycnal slope across the ACC despite increasing westerly winds have been taken as evidence that the ACC is close to an eddy-saturation regime<sup>19</sup>, as also seen in eddy-resolving numerical simulations<sup>20–22</sup>. Eddy compensation refers to the tendency for eddy mass transports to counter the wind-driven overturning circulation. In the limit of complete eddy compensation, the increase in northward Ekman transport in response to stronger winds is balanced by increased southward eddy mass transport. Eddy-resolving numerical simulations suggest that eddies only partially compensate the wind-driven circulation<sup>20–22</sup>. In addition, eddy fluxes and Ekman transport act at different depths and transport different water masses<sup>23</sup>.

In summary, substantial progress in the past two decades established a conceptual framework for the Southern Ocean, in which both wind and buoyancy forcing drive the circulation, the ACC and overturning are dynamically intertwined, eddies have a key role in establishing zonal-average dynamical balances, and interaction of flow with the sea floor balances forcing at the sea surface. However, this model fell short of a predictive theory for the response of the Southern Ocean to changes in forcing.

### Zonal asymmetry and regional dynamics

The zonal-average perspective illuminated many aspects of Southern Ocean dynamics, including how eddies provide a dynamical connection between the ACC and the overturning circulation. However, the Southern Ocean is not zonally uniform, and these asymmetries provide clues to missing physics. For example, although eddies are central to Southern Ocean dynamics, the distribution of eddy kinetic energy is not uniform, with relatively low levels over flat abyssal plains and elevated levels downstream of topography<sup>24</sup>. Recent studies have shown how many of the key physical processes relevant to Southern Ocean dynamics and climate are focused in ‘hot spots’, which are often linked to bottom topography. Figure 2 provides a schematic overview of the processes at work when the ACC encounters a topographic obstacle.

Upstream of the topography, the flow is largely zonal, eddy kinetic energy and eddy fluxes are low, vertical motion and cross-front exchange are suppressed, and the ACC fronts are distinct<sup>13</sup>. As the ACC



## Box 1

## Glossary

**Baroclinic and barotropic** Baroclinic flows vary with depth; barotropic flows are independent of depth.

**Baroclinic instability** A baroclinic instability is an instability of an atmospheric or oceanic flow that releases potential energy from the mean flow and increases the energy of the eddy field. The more rapidly the flow varies with depth (or, equivalently, the steeper the slope of isopycnals across the current), the more unstable the flow.

**Bottom-form stress** Bottom-form stress is a stress that is exerted on the sea floor by the flow, proportional to the difference in pressure at constant depth on either side of a topographic feature.

**Bottom-pressure torque** Bottom-pressure torque is a torque that is exerted on the sea floor by the flow, proportional to the difference in pressure along topography, at constant depth.

**Buoyancy** Buoyancy is added to the ocean by heating or by the input of fresh water by precipitation or ice melt; it is removed by cooling or by removal of fresh water by evaporation or the formation of sea ice.

**Diapycnal and isopycnal** Diapycnal refers to the direction perpendicular to surfaces of constant density (that is, isopycnals); isopycnal processes act along surfaces of constant density.

**Eddy** Ocean eddies are deviations from the mean flow, where the mean field can be defined by a temporal average (transient eddies) or a spatial average (stationary eddies). Eddies mostly transport properties along isopycnals. Baroclinic eddies are produced by baroclinic instability of the mean flow.

**Ekman** The Ekman layer is the surface layer of the ocean that is forced directly by the wind. In the Ekman layer, the drag force that is exerted by the wind stress is balanced by the pressure-gradient force and the Coriolis force. The Ekman transport is at right angles and to the left of the wind in the Southern Hemisphere, looking down-wind. Horizontal gradients in wind stress result in divergence or convergence of Ekman transport and hence upwelling or downwelling.

**Forcing** Ocean circulation is driven by the stress of the wind blowing on the sea surface and by factors that affect the buoyancy of the surface ocean (see 'buoyancy').

**Geostrophic balance** Large-scale flows in the ocean and atmosphere are close to geostrophic balance, whereby the pressure-gradient force is balanced by the Coriolis force.

**Kelvin wave** A Kelvin wave is a low-frequency gravity wave that is trapped to a land boundary or to the Equator.

**Meridional and zonal** Meridional refers to the north–south direction; zonal refers to the east–west direction.

**Planetary waves** Planetary waves or Rossby waves are wave motions that result from the conservation of potential vorticity and the fact that the Coriolis effect varies with latitude. Planetary waves have westward phase velocity and transfer information about changes in forcing.

**Potential vorticity** Vorticity refers to the rotation of a fluid element and includes contributions from horizontal gradients of velocity (relative vorticity) and from the Earth's rotation (planetary vorticity). Potential vorticity tends to be conserved by oceanographic flows, in the absence of dissipation.

**Topography** Topography is the bathymetry, or varying depth, of the sea floor.

**Tracer** A tracer is a generic property (namely, temperature or salinity) that is transported by ocean currents and mixing processes.

torque turns the flow equatorward and drives upwelling; where the deep flow crosses from shallow to deep, currents are turned poleward and associated with downwelling<sup>9,11,25</sup>. (Non-zero bottom-pressure torque implies a pressure difference along isobaths and hence a geostrophic flow towards or away from the boundary, which must be balanced by upslope or downslope flow<sup>9</sup>.) Pressure differences across topographic obstacles create stresses that, in the circumpolar integral, balance the wind stress at the sea surface<sup>14</sup>. Similarly, bottom-pressure torques balance the vorticity supplied by the curl of the wind stress, when integrated around the Southern Ocean. But the zonally integrated balances conceal the highly non-uniform distribution of bottom-form stress and bottom-pressure torque, which are concentrated where the ACC interacts with topography<sup>26,27</sup>. Eddy vorticity fluxes exert torques that turn the jets and help the ACC to navigate complex topography<sup>28</sup>. Topographic steering often causes jets to converge near topography, steepening isopycnals<sup>29</sup>. Although steeper isopycnals are more prone to baroclinic instability, sloping bathymetry provides a stabilizing influence because of the tendency to conserve potential vorticity and hence flow along, rather than across, bathymetric contours. Where deep gaps provide a pathway through ridges or other bathymetric obstacles, the ACC jets converge to pass through the gap, with eddies accelerating the deep flow (that is, the flow becomes more barotropic)<sup>30,31</sup>.

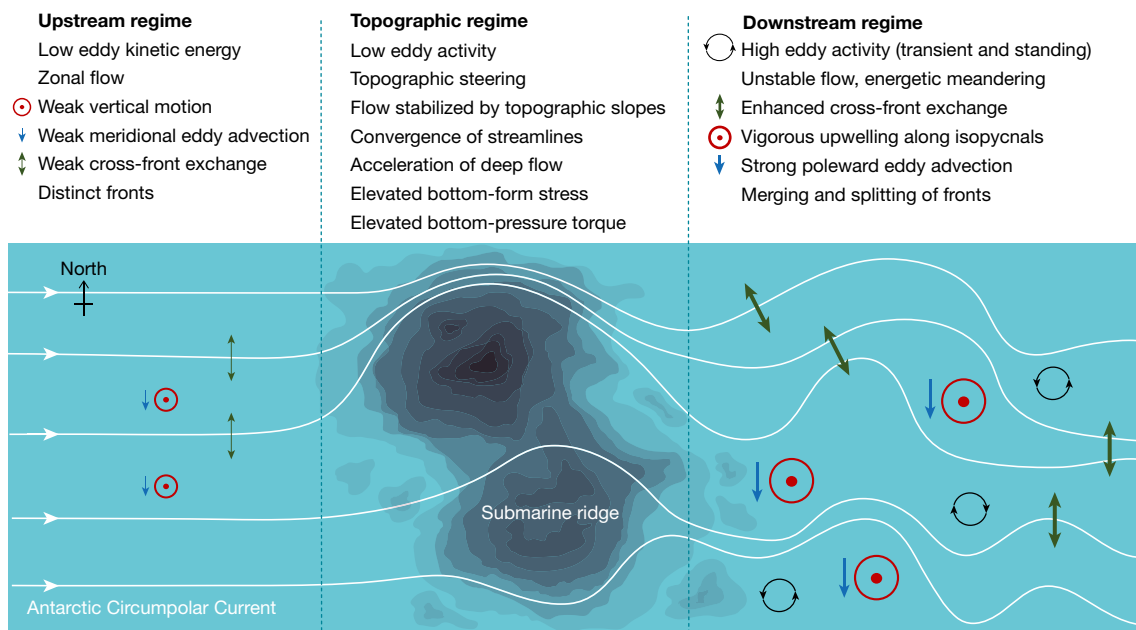
Downstream of the topography, the flow is no longer stabilized by topographic slopes and the potential energy stored in the topographically steepened isopycnals is released as a result of baroclinic instability<sup>32</sup>. Deviations from purely zonal flow by topographic steering help to destabilize the flow<sup>29,32</sup>. Eddy growth rates are high immediately downstream of the topography, whereas eddy kinetic energy reaches a maximum further downstream, after the eddies have had time to grow<sup>33</sup>. Meandering is often pronounced in the lee of topography and jets may merge and split<sup>29</sup>. Downstream of the ridge, advection and stirring by eddies are enhanced<sup>34</sup> and deep barotropic eddies facilitate cross-front exchange by increasing the angle between deep and shallow flows<sup>35</sup>.

Localized dynamics and deviations from zonal symmetry are also important for the overturning circulation. Whereas wind-driven upwelling into the surface mixed layer occurs over broad areas, model studies suggest that the poleward and upward motion of deep water in the ocean interior is focused in narrow regions where eddies and bottom topography facilitate vertical motion<sup>36,37</sup>. The circulation of deep water in the Southern Ocean interior therefore follows an upward spiral with mostly zonal flow along isopycnals and along depth horizons in regions of smooth topography, connected by rising and poleward flow where the ACC interacts with topographic obstacles and generates enhanced eddy fluxes. Upwelling of deep water occurs mainly along isopycnals; although weak diapycnal mixing over broad scales contributes to changes in density, most of the water-mass transformation occurs in the surface layer, where air–sea forcing and diapycnal mixing are both strong<sup>38</sup>. Vertical motion is also associated with the bottom-pressure torque that is created when the deep flow crosses isobaths<sup>9</sup>.

The subduction of mode and intermediate water by the upper cell of the Southern Ocean overturning circulation is also focused in local hot spots<sup>39</sup>. Subduction is large where horizontal flow crosses the sloping base of the surface mixed layer (a process known as lateral induction). Because the flow is steered by topography, and the spatial distribution of mixed layer depth is influenced by the large-scale circulation, the distribution of subduction is also influenced by bathymetry. The subduction of anthropogenic carbon into the interior is similarly focused in local hot spots<sup>40</sup>, including standing meanders of the ACC<sup>41</sup>.

Eddies also mix and stir tracers, primarily along isopycnals. It has been traditional in ocean modelling to assume mixing coefficients that are constant in time and space. Advances in theory, motivated by measurements of turbulence and tracer dispersion, have revealed both spatial and temporal variability in the strength of isopycnal mixing<sup>34,42</sup> (Fig. 3). Stirring along isopycnals is suppressed by one to two orders of magnitude in the core of the ACC jets, where the flow is sufficiently strong to carry tracers downstream before eddies have a chance to mix

encounters topography, stretching or squashing of the water column generates vorticity that is balanced by meridional motion (a consequence of the tendency of the flow to conserve angular momentum). Where the flow crosses isobaths from deep to shallow, bottom-pressure



**Fig. 2 | Interaction of the ACC with topography.** The schematic illustrates the dynamical processes at work in the distinct dynamical

regimes upstream, over ('topographic') and downstream of a topographic obstacle (represented by the shaded contours) in the path of the current.

cross-stream<sup>34,43</sup>. At depth, where the flow is weaker, the jets no longer inhibit stirring along isopycnals.

Microstructure measurements and tracer dispersion experiments have also revealed how the strength of diapycnal mixing varies in the Southern Ocean<sup>44–46</sup>, with the Diapycnal and Isopycnal Mixing Experiment in the Southern Ocean (DIMES) being the most notable example. Dissipation is enhanced in the upper 1,000 m by downward propagation and breaking of near-inertial waves generated by strong wind forcing<sup>47,48</sup>. Diapycnal mixing is also enhanced above rough topography, where internal lee waves generated by the interaction of the flow with topography propagate upwards and break<sup>46,49–51</sup>. The large diapycnal mixing rates at depth in the Southern Ocean help to drive the deep overturning cell and exchange between deep layers<sup>44,52</sup>. Deep diapycnal mixing in the Indian and Pacific oceans also has a critical role in the deep overturning cell: abyssal waters exported to the northern basins are converted by diapycnal mixing to slightly less-dense deep waters that return to the Southern Ocean and feed the upwelling limb of the overturning circulation<sup>5,53</sup>. In other words, the global overturning cell associated with sinking of dense water in the North Atlantic is closed first by deep mixing in the deep Indian and Pacific oceans and then by upwelling and water-mass transformations in the surface layer in the Southern Ocean. In the ocean interior, away from the sea surface and from rough topography, diapycnal mixing in the Southern Ocean is weak, as found in the rest of the global ocean<sup>45</sup>.

### Southern Ocean uptake of heat and carbon

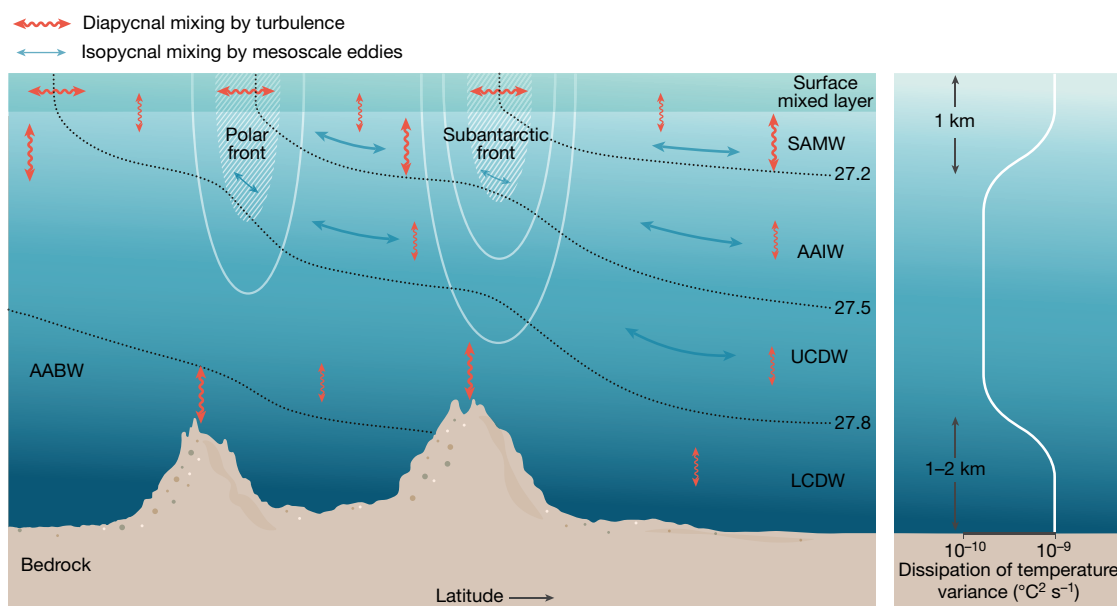
The strength of the Southern Ocean overturning circulation regulates the exchange of heat, carbon dioxide and other properties between the deep ocean and the surface layer and therefore has broad implications for climate and biogeochemical cycles. For example, the sinking of surface waters in the descending branches of the two overturning cells carries oxygen-rich water to ventilate the ocean interior. The upwelling limb of the overturning cells transfers nutrients from the deep ocean to the surface ocean, balancing the downward flux of nutrients and carbon via the sinking of organic matter; models suggest that upwelling and export of nutrients by the upper overturning cell support up to three-quarters of global marine primary production north of 30° S<sup>54,55</sup>. Likewise, the Southern Ocean influences atmospheric carbon dioxide levels on glacial–interglacial timescales: stronger upwelling of deep water vents more natural carbon to the atmosphere, warming the

climate during interglacial periods, whereas weaker upwelling results in more carbon being trapped in the deep ocean, cooling the climate during glacial periods<sup>56</sup>.

Of particular relevance to future climate is the role of the Southern Ocean overturning in the carbon and heat budget of the ocean. The ocean south of 40° S dominates the global ocean uptake of anthropogenic heat and carbon dioxide<sup>6,8,57,58</sup>. As the atmosphere warms and heats the ocean, the northward Ekman transport exports heat to the north, delaying warming south of the ACC and enhancing warming north of the ACC, where subduction of intermediate and mode waters carries heat into the ocean interior<sup>59</sup>. Over the past decade, the Southern Hemisphere has made the dominant contribution to the increase in global ocean heat content, reflecting both the transfer of heat to mode and intermediate waters by the overturning circulation and the deepening and spin-up of the subtropical gyres<sup>60,61</sup>. The upper cell of the overturning takes up and exports anthropogenic carbon dioxide in a similar manner<sup>7,57</sup>. For example, in a coupled climate–carbon model, the ocean south of 30° S (which represents 30% of the surface area of the global ocean) accounts for 75% ± 22% of anthropogenic heat uptake and 43% ± 3% of anthropogenic carbon dioxide uptake by the global ocean over the historical period<sup>8</sup>. The net exchange of carbon between the atmosphere and the Southern Ocean depends on two competing effects, both of which are influenced strongly by the overturning circulation: the outgassing of natural carbon driven by upwelling of carbon-rich deep water, and the uptake, transport and storage of anthropogenic carbon.

Given the prominent role of the Southern Ocean in the exchange of carbon between the atmosphere and the ocean, changes in the ability of the region to take up carbon dioxide would have substantial consequences for the global carbon cycle and climate. A decade ago, ocean models and atmospheric inversions suggested that the Southern Ocean carbon sink was 'saturated' and no longer keeping pace with increases in atmospheric carbon dioxide<sup>62,63</sup>. The reduction in the strength of the Southern Ocean carbon sink was attributed to increased outgassing of natural carbon associated with a wind-driven strengthening of the overturning circulation. This raised concerns that further weakening of the Southern Ocean carbon sink could contribute a positive feedback to climate change. At that time, there were insufficient ocean carbon data to estimate changes in ocean carbon uptake directly.





**Fig. 3 | Spatial distribution of mixing along and across isopycnals.** The schematic illustrates the spatial variability of mixing processes in the Southern Ocean. Isopycnal mixing (double-headed blue arrows) is inhibited in the upper part of the ACC jets (the Polar Front and Subantarctic Front; hatched region and contours indicate flow speed), where the mean flow is strong relative to the eddies (hatched region and contours, indicating flow speed). The strength of isopycnal mixing depends on eddy kinetic energy and hence wind strength. Diapycnal mixing (squiggly red arrows) is enhanced near the surface, where wind-generated, downward-propagating, inertial waves break, and near

topography, where lee waves generated by the interaction of the flow with topography propagate upwards and break. Diapycnal mixing is weak in the remainder of the domain. The dotted lines indicate contours of neutral density. The right panel shows the distribution of cross-isopycnal mixing (dissipation of temperature variance) with depth. The major water masses are labelled (AABW, Antarctic Bottom Water; LCDW, Lower Circumpolar Deep Water; UCDW, Upper Circumpolar Deep Water; AAIW, Antarctic Intermediate Water; SAMW, Subantarctic Mode Water). The right panel shows that cross-isopycnal mixing (dissipation of temperature variance) is typically enhanced in the upper ocean and within 1–2 km of the sea floor.

Longer and more complete time series of the partial pressure of carbon dioxide and new approaches to data analysis and mapping have revealed unanticipated variability in the Southern Ocean carbon sink<sup>64,65</sup>. As found in the earlier studies, the ocean carbon sink south of 35° S weakened in the 1990s, but strengthened again between 2002 and 2011 by 0.6 petagrams of carbon per year, or half the magnitude of the global trend in the ocean carbon sink over this period<sup>64</sup>. The ‘reinvigoration’ of the Southern Ocean carbon sink was attributed to changes in both the wind-driven overturning circulation<sup>64,65</sup> and the sea surface temperature (and hence the solubility of carbon dioxide), with colder temperatures dominating in the Pacific and weaker upwelling dominating in the Atlantic<sup>64</sup>. Whereas earlier studies emphasized the impact of changes in zonally averaged westerly winds on the Southern Ocean carbon sink, the more recent works show that the sink is also sensitive to regional wind anomalies. The large magnitude of temporal changes in the carbon sink underscores the importance of the Southern Ocean to global budgets and the sensitivity of this sink to temporal and regional variations in climate forcing.

### Sea-ice conveyor of fresh water

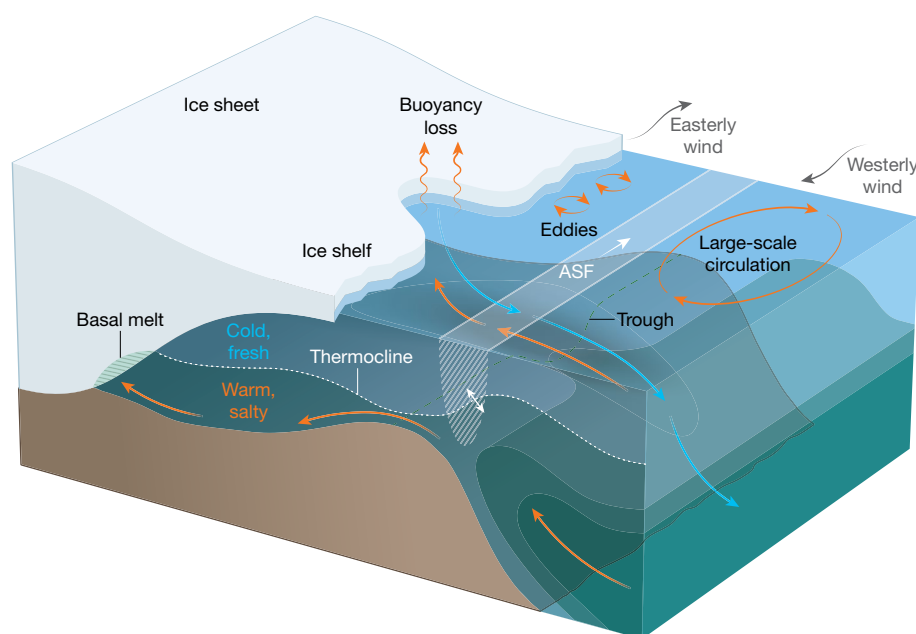
The existence of an overturning circulation in the Southern Ocean requires buoyancy forcing to transform water from one density class to another. More specifically, the conversion of upwelled deep water to lighter mode and intermediate waters requires an input of buoyancy through heating or addition of fresh water, whereas conversion to dense bottom water requires cooling or addition of salt during sea-ice formation. Most previous work has focused on buoyancy forcing by air–sea heat exchange and fresh water added by an excess of precipitation over evaporation<sup>5,66</sup>. However, improved estimates of sea-ice transport from observations and models indicate that ‘distillation’ of fresh water by sea-ice formation, transport and melt constitutes the dominant contribution to the buoyancy that is needed to close the upper cell<sup>67–70</sup>. Sea-ice formation rates are highest at high latitude, on the Antarctic continental shelf, where brine release contributes to the formation of dense shelf

water. Once formed, sea ice is exported to the north by Ekman transport and ocean gyres and melts. This results in a sink of fresh water over the continental shelf and a source of fresh water at lower latitudes. In this way, sea ice contributes to both the lower cell, where brine released during sea-ice formation contributes to the production of AABW, and the upper cell, where fresh water from sea-ice melt helps to convert deep water to lighter mode and intermediate water.

Because the strength of the overturning circulation is set by the buoyancy forcing<sup>3,17</sup>, changes in sea-ice formation and export may drive changes in the overturning circulation. Large regional trends in Antarctic sea-ice extent and persistence have been observed in recent decades<sup>68,71,72</sup>, with the retreat of ice in the eastern Pacific and growth of ice in the Ross Sea sector attributed to zonal asymmetries in wind forcing<sup>71</sup>; however, the response of the overturning circulation to the resulting regional anomalies in buoyancy forcing has not been investigated. Changes in fresh-water input from increases in high-latitude precipitation as the atmosphere warms<sup>73</sup> or from increased glacial melt<sup>74,75</sup> would also alter buoyancy forcing and water-mass formation in the Southern Ocean, with implications for the overturning circulation. The overturning circulation may also affect the distribution of sea ice, with the sign of the response depending on the timescale considered<sup>76</sup>. The instantaneous response to an increase in the westerly winds is the stronger northward Ekman transport of cold water and the expansion of sea ice. But continued strong upwelling eventually brings the deeper warm water to the surface, driving warming and sea-ice retreat. The short-timescale effect may help to explain the overall expansion of sea ice in recent decades, but cannot explain the zonal asymmetry between the sea-ice changes observed in the eastern and western Pacific.

### Ocean influence on the Antarctic Ice Sheet

Floating ice shelves buttress the Antarctic Ice Sheet by providing a back stress that resists the flow of glacial ice to the sea<sup>77</sup>. Thinning or retreat of ice shelves may therefore reduce the buttressing effect and cause increased export of ice and a rise in sea level<sup>78</sup>. Melt of the ice shelf



**Fig. 4 | Processes that control ocean heat flux to the Antarctic margin.**

The schematic illustrates physical processes that can influence the transport of warm water from the open ocean to the base of the floating ice shelves. The Antarctic Slope Front (ASF; white hatching and shading on the sea surface) forms the boundary between cold waters over the continental shelf and warm waters offshore. Processes that modify the strength and position of the Antarctic Slope Front (such as tides, wind or Kelvin waves) can influence the transport of warm water onto the shelf. Wind or buoyancy forcing (local or remote) can change the depth of the thermocline on the shelf (dotted white line), affecting how much warm water can reach the ice-shelf cavity. Eddies contribute to transferring

warm water across the edge of the continental shelf. Warm water can also be steered towards the shelf through deep troughs in the continental shelf. Heat loss in coastal polynyas can vent heat from warm waters on the shelf before they reach the ice shelves, thereby inhibiting basal melt, or drive export of cold, dense shelf waters (blue arrows) that is balanced by onshore transport of warm offshore waters (red arrows). The large-scale circulation (for example, gyres and the ACC) and changes in deep water properties or upwelling strength can influence the reservoir of ocean heat available for transport onto the shelf. Changes in wind can affect the upwelling of warm deep water and ocean currents so as to either facilitate or inhibit cross-shelf exchange of warm water.

from below by warm waters entering the sub-ice-shelf cavity influences the thickness of ice shelves and their buttressing capacity<sup>79</sup>. The parts of the ice sheet that are grounded below present-day sea level (that is, marine-based ice sheets) are particularly sensitive to ocean-driven change in the ice shelves at their seaward edge. The marine-based ice shelves in the Amundsen and Bellingshausen seas are more exposed to ocean heat flux than are other parts of the Antarctic margin because the warm waters of the ACC abut the continental shelf in that region. The most rapid thinning and mass loss has occurred in the Amundsen and Bellingshausen seas, where waters over the continental shelf have warmed<sup>80</sup>.

However, there is growing evidence that the marine-based portion of the East Antarctic Ice Sheet may also be vulnerable to ocean-driven melt. Satellite measurements show that parts of the East Antarctic Ice Sheet, including the Totten glacier that holds a volume of ice equivalent to more than 3.5 m of global sea-level rise, have thinned and that grounding lines have retreated in recent decades<sup>81</sup>. The Totten ice shelf experiences basal melt rates that are exceeded only by those of ice shelves in the Amundsen Sea<sup>82,83</sup>—a surprising result given the expectation that this part of the Antarctic margin was more isolated from warm ocean waters. But recent oceanographic observations on the continental shelf near the Totten ice shelf found that warm water was widespread, persisted through the year and reached the sub-ice-shelf cavity through a deep channel<sup>84,85</sup>. Recent simulations of the response of the ice sheet to future changes in climate suggest that the Aurora and Wilkes basins in East Antarctica will make a substantial contribution to future sea-level rise, with ice-sheet retreat initiated by ocean heat flux<sup>86,87</sup>.

To assess the potential vulnerability of the Antarctic Ice Sheet, the processes that regulate ocean heat transport to the sub-ice-shelf cavities, and their sensitivity to changes in forcing, need to be understood (Fig. 4). A key factor is the reservoir of ocean heat available for transfer across the shelf break, which is influenced by large-scale circulation

features such as the ACC, the Weddell and Ross gyres and wind-driven upwelling. The boundary between warm offshore waters and cooler waters over the continental shelf (the Antarctic Slope Front) can move or change in strength in response to local and remote forcing<sup>88</sup>, altering the temperature of water available for transport across the shelf break. Warm water present at the shelf break can be transported onto the shelf by eddies<sup>89</sup>, Kelvin waves<sup>88</sup> or by currents flowing along bathymetric contours in deep troughs on the continental shelf<sup>90</sup>. Air–sea interaction over the continental shelf influences how much of the heat that reaches the continental shelf makes it as far as the ice-shelf cavity. Strong ocean heat loss in polynyas can vent heat from the ocean before it reaches the ice shelf<sup>91</sup>. Wind forcing (local or remote) and polynya activity can alter the depth of the thermocline and so restrict or enhance the ocean heat flux to the cavity<sup>92</sup>. While ocean heat flux drives the melting of ice shelves, the input of glacial meltwater influences ocean circulation and sea ice<sup>74,93</sup>. Fresh water supplied by glacial melt increases the stratification of shelf waters, inhibiting deep convection and thereby reducing the formation of AABW and further enhancing basal melt by allowing ocean heat at depth to reach the ice shelves rather than be lost to the atmosphere<sup>75,94</sup>. Although progress has been made in identifying the processes that regulate ocean heat transport to ice-shelf cavities, it is not yet possible to determine the relative importance of these processes, now and in the future.

### Past and future change in the Southern Ocean

Given the effect of Southern Ocean processes on the global ocean circulation, climate and sea level, changes in the region could have widespread consequences. For example, changes in the amount of heat and carbon dioxide sequestered by the overturning circulation would act as a feedback on the rate of climate change. Increased ocean heat transport to ice-shelf cavities would drive increased basal melt, reduced buttressing, loss of mass from the Antarctic Ice Sheet and a rise in sea level<sup>86</sup>. Despite recent progress, understanding of Southern Ocean dynamics



still falls short of a complete theory that enables quantitative predictions of future changes in circulation. Nevertheless, recent observations of variability and change and advances in physical understanding provide some clues to guide a qualitative assessment of how the region will respond to changes in forcing.

Changes in various Southern Ocean properties have been documented in recent decades. The upper 2,000 m of the Southern Ocean has warmed and freshened<sup>19,73,95</sup> (such as by more than 0.1 °C per decade and more than 0.015 practical salinity units per decade at depths of 300–500 m over the past four decades<sup>19</sup>), with the largest changes in ocean heat content on the northern side of the ACC<sup>60</sup>. Both air–sea exchange and shifts in the position of the ACC have probably contributed to the changes in water properties<sup>96</sup>. Waters near the sea floor on the continental shelf of the Amundsen and Bellingshausen seas have warmed<sup>80</sup>. Eddy kinetic energy in the ACC has increased between the 1990s and the present<sup>97</sup>. The inventory of anthropogenic carbon has increased, with the largest changes in intermediate and mode waters north of the ACC<sup>6,58</sup>. Changes in chlorofluorocarbons between the 1990s and early 2000s have been interpreted as evidence for stronger upwelling of poorly ventilated deep water and stronger subduction of well-ventilated intermediate waters<sup>98</sup>, whereas more recent work suggests a reduction in overturning in the most recent decade<sup>65</sup>. Widespread freshening, warming and contraction of AABW has been observed over the past 30–50 years<sup>99–103</sup>.

What do these recent changes tell us about the sensitivity of the Southern Ocean to changes in forcing? Many of the changes summarized above are consistent with a spin-up of the wind-driven overturning cell. The westerly winds shifted south and strengthened between the 1980s and early 2000s, associated with a positive trend in the Southern Annular Mode, the dominant mode of variability of the Southern Hemisphere atmosphere<sup>104</sup>. Changes in wind forcing over the Southern Ocean have been linked to loss of ozone<sup>104</sup>, greenhouse gas forcing<sup>105</sup> and teleconnections to the tropics<sup>106</sup>. The changes in the ocean inventory of heat, fresh water and dissolved gases observed in recent decades are generally consistent with a strengthening of the wind-driven overturning, supporting the hypothesis that the eddy-driven circulation only partially compensates wind-driven changes in overturning<sup>107</sup>. Evidence from observations<sup>19</sup> and model studies<sup>20–22</sup> suggest that the ACC is close to the eddy saturation limit; hence, increases in wind forcing drive an increase in eddy kinetic energy, but little change in transport. Observations that demonstrate that the baroclinic transport of the ACC has remained roughly constant as wind forcing has strengthened<sup>19</sup>, while eddy kinetic energy has increased<sup>97</sup>, support this hypothesis.

How will the Southern Ocean change in the future? The westerly winds are expected to continue to strengthen and shift south in response to greenhouse gas forcing<sup>105</sup>. As the climate warms, we can anticipate an increase in heat input to the ocean, an increase in fresh-water input from enhanced precipitation<sup>73</sup> and ice melt<sup>68</sup> (including sea ice, icebergs and glacial melt), and hence an increase in the input of buoyancy to the ocean. The response of the Southern Ocean to these projected changes in forcing can be assessed by considering the contribution of different terms in the tracer balances.

Models and observations suggest that the first-order response to these changes in forcing will be passive advection of climate anomalies by the mean flow (that is,  $V\Delta T$ , where  $V$  is the unperturbed, three-dimensional, mean flow and  $\Delta T$  is the anomaly in tracer concentration)<sup>59,108</sup>. As the climate changes, the surface ocean is warmed, freshened and enriched in anthropogenic carbon dioxide by exchange with the atmosphere. These anomalies are swept north by the upper cell of the overturning circulation, increasing the inventory of anthropogenic heat, fresh water and carbon dioxide north of the ACC. The climate anomalies enter the interior ocean through localized subduction hot spots<sup>40,41</sup>; that is, anomalies in heat and other properties are harvested over broad spatial scales that are set by the wind and pumped into the interior through narrow windows whose distribution reflects interaction of the mean flow with the topography of the sea floor and of the mixed layer.

On the basis of the preceding discussion of Southern Ocean dynamics, we can anticipate that climate change will also alter the circulation, contributing to changes in water properties and inventories ( $\Delta VT_{\text{ave}}$ , where  $\Delta V$  is the climate-driven anomaly in circulation and  $T_{\text{ave}}$  is the mean tracer concentration). Stronger winds will drive a stronger wind-driven cell, which will be partially compensated by a more energetic eddy-driven cell<sup>21</sup>. Warming and increased fresh-water input will increase the buoyancy input to the ocean, driving the stronger water-mass transformations required by a spin-up of the upper cell of the overturning circulation. The buoyancy transport by the overturning circulation must increase to balance the enhanced buoyancy input, through an increase in the strength of the overturning circulation<sup>3,17</sup> or in the density difference between the upper and lower branches of the overturning (as illustrated in an idealized model<sup>23</sup>). An increase in strength of the upper cell will act on mean isopycnal and diapycnal gradients to transport heat and other climate properties. Because eddy diffusivity is a function of eddy kinetic energy, an increase in eddy energy as a consequence of increased wind forcing will strengthen eddy transport along isopycnals<sup>107</sup>. A stronger upper cell will drive changes in the subsurface ocean by increasing the poleward and upward transport of old, poorly ventilated deep water south of the ACC and the subduction of young, well-ventilated mode and intermediate waters north of the ACC<sup>108</sup>. This spin-up of the overturning circulation acts in the same sense as passive advection of climate anomalies by the mean flow, flushing climate anomalies from south to north across the ACC and increasing the inventory in subducted mode and intermediate waters. Therefore, we expect warming, freshening and increased storage of anthropogenic carbon dioxide north of the ACC and smaller changes in inventories south of the ACC.

The future of the Southern Ocean carbon sink is difficult to assess: stronger upwelling will mean more outgassing of natural carbon dioxide, whereas a stronger upper cell will take up and store more anthropogenic carbon dioxide<sup>57,58,63</sup>. Whether the net Southern Ocean carbon sink increases or decreases depends on the extent to which eddies compensate changes in the wind-driven overturning, and the response of the eddy-driven cell remains uncertain. Changes in surface temperature in response to regional and larger-scale wind anomalies will also influence the strength of the carbon sink<sup>64</sup>. Stronger upwelling of relatively warm deep water may also increase the ocean heat transport to the base of floating ice shelves; but, as outlined above, the delivery of heat to the ice-shelf cavities depends on many factors which themselves are likely to be affected by changes in climate forcing.

Changes in surface forcing will drive further changes in the Southern Ocean that will have consequences for climate. Warming and freshening as a result of air–sea exchange and ice melt will enhance stratification in the upper ocean, inhibiting exchange between the mixed layer and the interior<sup>69</sup>, weakening deep convection<sup>109</sup> and reducing the density of shelf waters that contribute to AABW formation<sup>99</sup>. We might therefore anticipate a weakening of the lower cell, although changes in recent decades suggest lighter bottom waters will continue to ventilate the abyssal ocean<sup>103</sup> until a threshold is reached when winter shelf waters are too light to sink to the abyssal ocean. Increased winds may drive increased diapycnal mixing at depth over rough topography<sup>46</sup>, increasing a ‘short circuit’<sup>44</sup> in the abyssal ocean that would reduce the amount of well-ventilated bottom water that reaches the basins to the north, further reducing the influence of the lower cell.

A consistent theme of this Review has been the importance of local and regional dynamics, which are often linked to topography. Although the topography does not change with time, and we can therefore anticipate that the same topographic features will continue to localize the dynamics of the Southern Ocean circulation, changes in the path of the current will affect dynamical balances. For example, small changes in the angle at which the flow intersects topography can change the torque that is exerted by the flow on the sea floor<sup>27</sup>. Likewise, changes in the path of the ACC will alter pressure differences across topography, changing the stress that is exerted by the flow on the sea floor<sup>26</sup>. Standing meanders in the lee of topography are probably of particular

importance, where stronger wind forcing drives an increase in mean-dering, increased instability, more eddy activity, enhanced downward transfer of momentum and acceleration of the deep flow; interaction of the deep flow with topography establishes bottom-form stress and bottom-pressure torque to balance the forcing<sup>32</sup>. Analysis of the time-dependent momentum balance of the ACC suggests that the adjustment to a change in wind involves rapid barotropic processes that enable a nearly instantaneous response of bottom-form stress to changes in wind forcing<sup>26</sup>.

The above discussion of the future of the Southern Ocean is informed by recent progress in dynamical understanding, but remains speculative, reflecting both gaps in the theoretical underpinning of the Southern Ocean circulation and uncertainties in future climate forcing. The most substantial gap in physical understanding is the response of the eddy field to changes in forcing. This Review has highlighted how eddies—transient and stationary—are intimately involved in almost every aspect of Southern Ocean dynamics, helping to set the strength and vertical and horizontal structure of the mean flow (the ACC and the overturning), to drive meridional flow, to navigate complex topography, to shape the complex temporal and spatial distribution of ocean mixing, and to transport energy, momentum, vorticity and tracers.

## Outlook and open questions

Substantial progress has been made in recent years in understanding the dynamics and global influence of the Southern Ocean, underpinned by a revolution in ocean observing and advances in theory and numerical simulation. These discoveries have shown that the Southern Ocean needs to be viewed through both wide-angle and close-up lenses. Southern Ocean processes have a disproportionate effect on global climate, biogeochemical cycles and sea level, and are linked to low latitudes through diverse teleconnections that involve interactions between the atmosphere, ocean and cryosphere. On the other hand, these global effects are the expression of dynamics that play out largely on local and regional scales, often mediated by topography.

Many questions remain open. Perhaps of greatest importance to climate and sea level is the uncertain response of the Southern Ocean overturning circulation—in particular, the eddy-driven contribution—to changes in wind and buoyancy forcing. Eddies are now known to transfer momentum and vorticity from the sea surface to the sea floor, but the detailed pathways and their sensitivity to changes in forcing are unknown. Recent studies have highlighted how the upwelling and downwelling limbs of the overturning circulation are localized by topography, as is cross-front exchange, but the three-dimensional structure of the overturning remains obscure. Knowledge of the nature and causes of variability in the Southern Ocean is rudimentary, including the relative contributions of local forcing, teleconnections to low latitude and intrinsic variability. The theoretical foundation for Southern Ocean dynamics is developing rapidly, but remains incomplete. This gap is reflected, for example, by the speculative nature of the above discussion of the response of the Southern Ocean to changes in forcing and by our inability to do much more than list the mechanisms that influence the delivery of ocean heat to the Antarctic margin. The fact that new observations continue to reveal surprises that challenge existing thinking also underscores gaps in knowledge; examples include the unanticipated variability of the Southern Ocean carbon sink, the dominant contribution of the Southern Hemisphere to the change in global ocean heat content in the past decade, and evidence that the East Antarctic Ice Sheet is more exposed to ocean heat transport than once thought.

The prospects for further progress are encouraging. The Argo array of profiling floats has allowed year-round, broad-scale observations of the data-sparse Southern Ocean for the past decade, and many of the recent insights summarized in this Review have relied on these new measurements. As the profiling float array expands into the ice-covered and deep ocean, further breakthroughs will follow. The capability of other observing tools such as gliders and autonomous underwater vehicles is also increasing, making measurements feasible in previously inaccessible areas such as beneath floating ice shelves and perennial

ice. Continued growth in computing power is allowing numerical exploration of Southern Ocean dynamics with high-resolution models that resolve critical physical processes that act at small scales. Insights gained in this way promise to improve the parameterizations used in Earth system models, helping to reduce biases in their representation of the Southern Ocean. The combination of new observations, advances in theory and improvements in modelling promises to deliver a much better understanding of how the Southern Ocean circulation will respond to future change and how it will influence global climate, biogeochemical cycles and sea-level rise.

Received: 13 November 2017; Accepted: 27 March 2018;

Published online 13 June 2018.

1. Deacon, G. E. R. The hydrology of the Southern Ocean. *Discov. Rep.* **15**, 1–124 (1937).
2. Sverdrup, H. U. On vertical circulation in the ocean due to the action of the wind with application to conditions within the Antarctic Circumpolar Current. *Discov. Rep.* **VII**, 139–170 (1933).
3. Speer, K., Rintoul, S. R. & Sloyan, B. The diabatic Deacon cell. *J. Phys. Oceanogr.* **30**, 3212–3222 (2000).  
**This study highlights the role of the Southern Ocean in closing the global overturning circulation.**
4. Marshall, J. & Speer, K. Closure of the meridional overturning circulation through Southern Ocean upwelling. *Nat. Geosci.* **5**, 171–180 (2012).  
**This paper provides a review of the Southern Ocean overturning circulation and its role in the Earth system.**
5. Sloyan, B. M. & Rintoul, S. R. The Southern Ocean limb of the global deep overturning circulation. *J. Phys. Oceanogr.* **31**, 143–173 (2001).
6. Sabine, C. L. et al. The oceanic sink for anthropogenic CO<sub>2</sub>. *Science* **305**, 367–371 (2004).
7. Gruber, N. et al. Oceanic sources, sinks, and transport of atmospheric CO<sub>2</sub>. *Glob. Biogeochem. Cycles* **23**, GB1005 (2009).
8. Frölicher, T. L. et al. Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *J. Clim.* **28**, 862–886 (2015).  
**This study, which is based on numerical model simulations, demonstrates the dominant contribution of the Southern Ocean to the uptake of anthropogenic heat and carbon dioxide.**
9. Hughes, C. W. Nonlinear vorticity balance of the Antarctic Circumpolar Current. *J. Geophys. Res.* **110**, C11008 (2005).  
**This paper provides a lucid explanation of the vorticity balance in the Southern Ocean.**
10. Anderson, D. L. T. & Gill, A. E. Spin-up of a stratified ocean with applications to upwelling. *Deep Sea Res.* **22**, 583–596 (1975).
11. Rintoul, S. R., Hughes, C. & Olbers, D. in *Ocean Circulation and Climate* (eds Siedler, G. et al.) 271–302 (Academic Press, Cambridge, 2001).
12. Olbers, D., Borowski, D., Völker, C. & Wölff, J.-O. The dynamical balance, transport and circulation of the Antarctic Circumpolar Current. *Antarct. Sci.* **16**, 439–470 (2004).
13. Rintoul, S. R. & Naveira Garabato, A. C. in *Ocean Circulation and Climate* 2nd edn (eds Siedler, G. et al.) Ch. 18 (Academic Press, Cambridge, 2013).  
**This review of Southern Ocean dynamics provides additional detail on some of the processes highlighted here.**
14. Munk, W. H. & Palmén, E. Note on the dynamics of the Antarctic Circumpolar Current. *Tellus* **3**, 53–55 (1951).
15. Johnson, G. C. & Bryden, H. On the size of the Antarctic Circumpolar Current. *Deep Sea Res. Part A* **36**, 39–53 (1989).
16. Hogg, A. McC. An Antarctic Circumpolar Current driven by surface buoyancy forcing. *Geophys. Res. Lett.* **37**, L23601 (2010).
17. Marshall, J. & Radko, T. Residual-mean solutions for the Antarctic Circumpolar Current and its associated overturning circulation. *J. Phys. Oceanogr.* **33**, 2341–2354 (2003).
18. Straub, D. N. On the transport and angular momentum balance of channel models of the Antarctic Circumpolar Current. *J. Phys. Oceanogr.* **23**, 776–782 (1993).
19. Böning, C. W., Disper, A., Visbeck, M., Rintoul, S. R. & Schwarzkopf, F. Response of the Antarctic Circumpolar Current to recent climate change. *Nat. Geosci.* **1**, 864–869 (2008).
20. Hallberg, R. & Gnanadesikan, A. The role of eddies in determining the structure and response of the wind-driven Southern Hemisphere overturning: results from the modeling eddies in the Southern Ocean (MESO) project. *J. Phys. Oceanogr.* **36**, 2232–2252 (2006).
21. Farneti, R., Delworth, T. L., Rosati, A. J., Griffies, S. M. & Zeng, F. The role of mesoscale eddies in the rectification of the Southern Ocean response to climate change. *J. Phys. Oceanogr.* **40**, 1539–1557 (2010).
22. Dufour, C. O. et al. Standing and transient eddies in the response of the Southern Ocean meridional overturning to the Southern Annular Mode. *J. Clim.* **25**, 6958–6974 (2012).
23. Morrison, A. K. & Hogg, A. McC. On the relationship between Southern Ocean overturning and ACC transport. *J. Phys. Oceanogr.* **43**, 140–148 (2013).
24. Chelton, D. B., Schlax, M. G., Samelson, R. M. & deSzoeke, R. A. Global observations of large ocean eddies. *Geophys. Res. Lett.* **34**, L15606 (2007).



25. Sokolov, S. & Rintoul, S. R. On the relationship between fronts of the Antarctic Circumpolar Current and surface chlorophyll concentrations in the Southern Ocean. *J. Geophys. Res. Oceans* **112**, C07030 (2007).
26. Masich, J., Chereskin, T. K. & Mazloff, M. Topographic form stress in the Southern Ocean state estimate. *J. Geophys. Res.* **120**, 7919–7933 (2015).
27. Firing, Y. I., Chereskin, T. K., Watts, D. R. & Mazloff, M. R. Bottom pressure torque and the vorticity balance from observations in Drake Passage. *J. Geophys. Res. Oceans* **121**, 4282–4302 (2016).
28. Williams, R. G., Wilson, C. & Hughes, C. W. Ocean and atmosphere storm tracks: the role of eddy vorticity forcing. *J. Phys. Oceanogr.* **37**, 2267–2289 (2007).
29. Thompson, A. F. & Sallée, J. B. Jets and topography: jet transitions and the impact on transport in the Antarctic Circumpolar Current. *J. Phys. Oceanogr.* **42**, 956–972 (2012).
30. Smith, I. J., Stevens, D. P., Heywood, K. J. & Meredith, M. P. The flow of the Antarctic Circumpolar Current over the North Scotia Ridge. *Deep Sea Res. Part 1* **57**, 14–28 (2010).
31. Rintoul, S. R. et al. Antarctic Circumpolar Current transport and barotropic transition at Macquarie Ridge. *Geophys. Res. Lett.* **41**, 7254–7261 (2014).
32. Thompson, A. F. & Naveira Garabato, A. C. Equilibration of the Antarctic Circumpolar Current by standing meanders. *J. Phys. Oceanogr.* **44**, 1811–1828 (2014).
- This study shows how changes in the path of the Antarctic Circumpolar Current ('flexing' of meanders) can give rise to eddy-mean flow and flow-topography interactions that balance changes in forcing.**
33. Dufour, C. O. et al. Role of mesoscale eddies in cross-frontal transport of heat and biogeochemical tracers in the Southern Ocean. *J. Phys. Oceanogr.* **45**, 3057–3081 (2015).
34. Naveira Garabato, A. C., Ferrari, R. & Polzin, K. L. Eddy stirring in the Southern Ocean. *J. Geophys. Res.* **116**, C09019 (2011).
- This paper provides a comprehensive examination of along-isopycnal stirring in the Southern Ocean by eddies.**
35. Chereskin, T. K. et al. Strong bottom currents and cyclogenesis in Drake Passage. *Geophys. Res. Lett.* **36**, L23602 (2009).
36. Döös, K., Nycander, J. & Coward, A. C. Lagrangian decomposition of the Deacon Cell. *J. Geophys. Res.* **113**, C07028 (2008).
37. Tamsitt, V. et al. Spiraling pathways of global deep waters to the surface of the Southern Ocean. *Nat. Commun.* **8**, 172 (2017); corrigendum **9**, 209 (2018).
38. Tamsitt, V., Abernathey, R. P., Mazloff, M. R., Wang, J. & Talley, L. D. Transformation of deep water masses along Lagrangian upwelling pathways in the Southern Ocean. *J. Geophys. Res. Oceans* **123**, 1994–2017 (2018).
39. Sallée, J. B., Rintoul, S. R. & Wijffels, S. E. Southern ocean thermocline ventilation. *J. Phys. Oceanogr.* **40**, 509–529 (2010).
40. Sallée, J. B., Matear, R., Rintoul, S. R. & Lenton, A. Surface to interior pathways of anthropogenic CO<sub>2</sub> in the southern hemisphere oceans. *Nat. Geosci.* **5**, 579–584 (2012).
41. Langlais, C. E. et al. Stationary Rossby waves dominate subduction of anthropogenic carbon in the Southern Ocean. *Sci. Rep.* **7**, 17076 (2017).
42. Tulloch, R. et al. Direct estimate of lateral eddy diffusivity upstream of Drake Passage. *J. Phys. Oceanogr.* **44**, 2593–2616 (2014).
43. Ferrari, R. & Nikurashin, M. Suppression of eddy diffusivity across jets in the Southern Ocean. *J. Phys. Oceanogr.* **40**, 1501–1519 (2010).
- This study explains how the strong jets of the Antarctic Circumpolar Current suppress eddy stirring across the current.**
44. Garabato Naveira, A. C., Stevens, D. P., Watson, A. J. & Roether, W. Short-circuiting of the oceanic overturning circulation in the Antarctic Circumpolar Current. *Nature* **447**, 194–197 (2007).
45. Ledwell, J. R., St. Laurent, L. C., Giron, J. B. & Toole, J. M. Diapycnal mixing in the Antarctic Circumpolar Current. *J. Phys. Oceanogr.* **41**, 241–246 (2011).
46. Naveira Garabato, A. C., Polzin, K. L., Ferrari, R., Zika, J. D. & Forryan, A. A microscale view of mixing and overturning across the Antarctic Circumpolar Current. *J. Phys. Oceanogr.* **46**, 233–254 (2016).
47. Waterman, S. N., Naveira Garabato, A. C. & Polzin, K. L. Internal waves and turbulence in the Antarctic Circumpolar Current. *J. Phys. Oceanogr.* **43**, 259–282 (2013).
48. Sheen, K. L. et al. Rates and mechanisms of turbulent dissipation and mixing in the Southern Ocean: results from the Diapycnal and Isopycnal Mixing Experiment in the Southern Ocean (DIMES). *J. Geophys. Res. Oceans* **118**, 2774–2792 (2013).
49. Nikurashin, M. & Ferrari, R. Radiation and dissipation of internal waves generated by geostrophic motions impinging on small-scale topography: application to the Southern Ocean. *J. Phys. Oceanogr.* **40**, 2025–2042 (2010).
50. Laurent, L. St. et al. Turbulence and diapycnal mixing in Drake Passage. *J. Phys. Oceanogr.* **42**, 2143–2152 (2012).
51. Watson, A. J. et al. Rapid cross-density ocean mixing at mid-depths in the Drake Passage measured by tracer release. *Nature* **501**, 408–411 (2013).
- Using observations of the spreading of a tracer released in the Southern Ocean, the authors show that diapycnal mixing is rapid where the Antarctic Circumpolar Current interacts with rough topography.**
52. Nikurashin, M. & Ferrari, R. Overturning circulation driven by breaking internal waves in the deep ocean. *Geophys. Res. Lett.* **40**, 3133–3137 (2013).
53. Talley, L. D. Closure of the global overturning circulation through the Indian, Pacific, and Southern oceans: schematics and transports. *Oceanography* **26**, 80–97 (2013).
54. Sarmiento, J. L., Gruber, N., Brzezinski, M. A. & Dunne, J. P. High-latitude controls of thermocline nutrients and low latitude biological productivity. *Nature* **427**, 56–60 (2004); corrigendum **479**, 556 (2011).
55. Marinov, I., Gnanadesikan, A., Toggweiler, J. R. & Sarmiento, J. L. The Southern Ocean biogeochemical divide. *Nature* **441**, 964–967 (2006).
56. Sigman, D. M., Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in atmospheric CO<sub>2</sub> concentration. *Nature* **466**, 47–55 (2010).
57. Mikaloff Fletcher, S. E. et al. Inverse estimates of anthropogenic CO<sub>2</sub> uptake, transport, and storage by the ocean. *Glob. Biogeochem. Cycles* **20**, GB2002 (2006).
58. Khatala, S. et al. Global ocean storage of anthropogenic carbon. *Biogeosciences* **10**, 2169–2191 (2013).
59. Armour, K. C., Marshall, J., Scott, J. R., Donohoe, A. & Newsom, E. R. Southern Ocean warming delayed by circumpolar upwelling and equatorward transport. *Nat. Geosci.* **9**, 549–554 (2016).
60. Roemmich, D. J. et al. Unabated planetary warming and its ocean structure since 2006. *Nat. Clim. Change* **5**, 240–245 (2015).
61. Gao, L., Rintoul, S. R. & Yu, W. Recent wind-driven changes in Subantarctic Mode Water and its impact on ocean heat storage. *Nat. Clim. Change* **8**, 58–63 (2018).
62. Le Quééré, C. et al. Saturation of the southern ocean CO<sub>2</sub> sink due to recent climate change. *Science* **316**, 1735–1738 (2007).
63. Lovenduski, N. S., Gruber, N. & Doney, S. C. Toward a mechanistic understanding of the decadal trends in the Southern Ocean carbon sink. *Glob. Biogeochem. Cycles* **22**, GB3016 (2008).
64. Landschützer, P. et al. The reinvigoration of the Southern Ocean carbon sink. *Science* **349**, 1221–1224 (2015).
65. DeVries, T., Holzer, M. & Primeau, F. Recent increase in oceanic carbon uptake driven by weaker upper-ocean overturning. *Nature* **542**, 215–218 (2017).
66. Lumpkin, R. & Speer, K. Global ocean meridional overturning. *J. Phys. Oceanogr.* **37**, 2550–2562 (2007).
67. Abernathey, R. P. et al. Water-mass transformation by sea ice in the upper branch of the Southern Ocean overturning. *Nat. Geosci.* **9**, 596–601 (2016).
68. Haumann, F. A., Gruber, N., Münnich, M., Frenger, I. & Kern, S. Sea-ice transport driving Southern Ocean salinity and its recent trends. *Nature* **537**, 89–92 (2016).
- This study highlights the contribution of fresh-water transport by sea ice to the buoyancy budget and water-mass transformations that are central to the Southern Ocean overturning circulation.**
69. Pellichero, V., Sallée, J.-B., Schmidtko, S., Roquet, F. & Charrassin, J.-B. The ocean mixed layer under Southern Ocean sea-ice: seasonal cycle and forcing. *J. Geophys. Res. Oceans* **122**, 1608–1633 (2017).
70. Pellichero, V., Sallée, J.-B., Chapman, C. C. & Downes, S. M. The southern ocean meridional overturning in the sea-ice sector is driven by freshwater fluxes. *Nat. Commun.* **9**, 1789 (2018).
71. Holland, P. R. & Kwok, R. Wind-driven trends in Antarctic sea-ice drift. *Nat. Geosci.* **5**, 872–875 (2012).
72. Hobbs, W. R. et al. A review of recent changes in Southern Ocean sea ice, their drivers and forcings. *Global Planet. Change* **143**, 228–250 (2016).
73. Durack, P. J., Wijffels, S. E. & Matear, R. J. Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* **336**, 455–458 (2012).
74. Bintanja, R., van Oldenborgh, G. J., Drijfhout, S. S., Wouters, B. & Katsman, C. A. Important role for ocean warming and increased ice-shelf melt in Antarctic sea-ice expansion. *Nat. Geosci.* **6**, 376–379 (2013).
75. Silvano, A. et al. Freshening by glacial meltwater enhances melting of ice shelves and reduces formation of Antarctic Bottom Water. *Sci. Adv.* **4**, eaap9467 (2018).
76. Ferreira, D., Marshall, J., Bitz, C. M., Solomon, S. & Plumb, A. Antarctic ocean and sea ice response to ozone depletion: a two-time-scale problem. *J. Clim.* **28**, 1206–1226 (2015).
77. Shepherd, A., Fricker, H. A. & Farrell, S. L. Trends and connections across the Antarctic cryosphere. *Nature* **558**, <https://doi.org/10.1038/s41586-018-0171-6> (2018).
78. Dupont, T. K. & Alley, R. B. Assessment of the importance of ice-shelf buttressing to ice-sheet flow. *Geophys. Res. Lett.* **32**, L04503 (2005).
79. Pritchard, H. D. et al. Antarctic ice-sheet loss driven by basal melting of ice shelves. *Nature* **484**, 502–505 (2012).
80. Schmidtko, S., Heywood, K. J., Thompson, A. F. & Aoki, S. Multidecadal warming of Antarctic waters. *Science* **346**, 1227–1231 (2014).
81. Li, X., Rignot, E., Morlighem, M., Mougnot, J. & Scheuchl, B. Grounding line retreat of Totten Glacier, East Antarctica, 1996 to 2013. *Geophys. Res. Lett.* **42**, 8049–8056 (2015).
82. Rignot, E., Jacobs, S., Mougnot, J. & Scheuchl, B. Ice-shelf melting around Antarctica. *Science* **341**, 266–270 (2013).
83. Depoorter, M. A. et al. Calving fluxes and basal melt rates of Antarctic ice shelves. *Nature* **502**, 89–92 (2013).
84. Rintoul, S. R. et al. Ocean heat drives rapid basal melt of the Totten Ice Shelf. *Sci. Adv.* **2**, e1601610 (2016).
85. Silvano, A., Rintoul, S. R., Peña-Molino, B. & Williams, G. D. Distribution of water masses and meltwater on the continental shelf near the Totten and Moscow University ice shelves. *J. Geophys. Res. Oceans* **122**, 2050–2068 (2017).
86. Golledge, N. R. et al. The multi-millennial Antarctic commitment to future sea-level rise. *Nature* **526**, 421–425 (2015).
87. DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. *Nature* **531**, 591–597 (2016).
88. Spence, P. et al. Localized rapid warming of West Antarctic subsurface waters by remote winds. *Nat. Clim. Change* **7**, 595–603 (2017).

89. Stewart, A. L. & Thompson, A. F. Eddy-mediated transport of warm Circumpolar Deep Water across the Antarctic Shelf Break. *Geophys. Res. Lett.* **42**, 432–440 (2015).
90. Dinniman, M. S., Klinck, J. M. & Smith, W. O. Jr. A model study of Circumpolar Deep Water on the West Antarctic Peninsula and Ross Sea continental shelves. *Deep Sea Res. Part II* **58**, 1508–1523 (2011).
91. Khazendar, A. et al. Observed thinning of Totten Glacier is linked to coastal polynya variability. *Nat. Commun.* **4**, 2857 (2013).
92. Dutrieux, P. et al. Strong sensitivity of Pine Island ice-shelf melting to climatic variability. *Science* **343**, 174–178 (2014).
93. Pauling, A. G., Smith, I. J., Langhorne, P. J. & Bitz, C. M. Time-dependent freshwater input from ice shelves: impacts on Antarctic sea ice and the Southern Ocean in an Earth system model. *Geophys. Res. Lett.* **44**, 10454–10461 (2017).
94. Hellmer, H. H. Impact of Antarctic ice shelf basal melting on sea ice and deep ocean properties. *Geophys. Res. Lett.* **31**, L10307 (2004).
95. Gille, S. T. Decadal-scale temperature trends in the Southern Hemisphere ocean. *J. Clim.* **21**, 4749–4765 (2008).
96. Meijers, A. J. S., Bindoff, N. L. & Rintoul, S. R. Frontal movements and property fluxes: contributions to heat and freshwater trends in the Southern Ocean. *J. Geophys. Res. Oceans* **116**, C08024 (2011).
97. Hogg, A. McC. et al. Recent trends in the Southern Ocean eddy field. *J. Geophys. Res. Oceans* **120**, 257–267 (2015).
98. Waugh, D. W., Primeau, F., DeVries, T. & Holzer, M. Recent changes in the ventilation of the southern oceans. *Science* **339**, 568–570 (2012).
99. Jacobs, S. S. & Giulivi, C. F. Large multidecadal salinity trends near the Pacific-Antarctic continental margin. *J. Clim.* **23**, 4508–4524 (2010).
100. Purkey, S. G. & Johnson, G. C. Warming of global abyssal and deep southern ocean waters between the 1990s and 2000s: contributions to global heat and sea level rise budgets. *J. Clim.* **23**, 6336–6351 (2010).
101. Purkey, S. G. & Johnson, G. C. Global contraction of Antarctic Bottom Water between the 1980s and 2000s. *J. Clim.* **25**, 5830–5844 (2012).
102. Purkey, S. G. & Johnson, G. C. Antarctic Bottom Water warming and freshening: contributions to sea level rise, ocean freshwater budgets, and global heat gain. *J. Clim.* **26**, 6105–6122 (2013).
103. van Wijk, E. M. & Rintoul, S. R. Freshening drives contraction of Antarctic Bottom Water in the Australian Antarctic Basin. *Geophys. Res. Lett.* **41**, 1657–1664 (2014).
104. Thompson, D. W. J. et al. Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change. *Nat. Geosci.* **4**, 741–749 (2011).
105. Swart, N. C. & Fyfe, J. C. Observed and simulated changes in the Southern Hemisphere surface westerly wind-stress. *Geophys. Res. Lett.* **39**, L16711 (2012).
106. Ding, Q., Steig, E. J., Battisti, D. S. & Wallace, J. M. Influence of the tropics on the Southern Annular Mode. *J. Clim.* **25**, 6330–6348 (2012).
107. Meredith, M. P., Naveira Garabato, A. C., Hogg, A. McC. & Farneti, R. Sensitivity of the overturning circulation in the Southern Ocean to decadal changes in wind forcing. *J. Clim.* **25**, 99–110 (2012).
108. Morrison, A. K., Griffies, S. M., Winton, M., Anderson, W. G. & Sarmiento, J. L. Mechanisms of Southern Ocean heat uptake and transport in a global eddying climate model. *J. Clim.* **29**, 2059–2075 (2016).
109. Ito, T. et al. Sustained growth of the Southern Ocean carbon storage in a warming climate. *Geophys. Res. Lett.* **42**, 4516–4522 (2015).
110. Patara, L., Böning, C. W. B. & Biastoch, A. Variability and trends in Southern Ocean eddy activity in 1/12° ocean model simulations. *Geophys. Res. Lett.* **43**, 4517–4523 (2016).
111. Rintoul, S. R. Southern Ocean currents and climate. *Pap. Proc. R. Soc. Tasman.* **133**, 41–50 (2000).

**Acknowledgements** A. Silvano, A. Foppert, A. Lenton, M. Nikurashin and E. van Wijk provided comments on the paper. M. Bessel and G. Wells prepared the original version of Fig. 1b. This work is supported in part by the Australian Government Cooperative Research Centre (CRC) programme through the Antarctic Climate and Ecosystems CRC, by the National Environmental Science Program, by the Centre for Southern Hemisphere Oceans Research, a partnership between CSIRO and the Qingdao National Laboratory for Marine Science and Technology, and by the Tinker-Muse Prize for Science and Policy in Antarctica.

**Reviewer information** Nature thanks R. Ferrari, N. Gruber and K. Speer for their contribution to the peer review of this work.

**Competing interests** The author declares no competing interests.

#### Additional information

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to S.R.R.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Mass balance of the Antarctic Ice Sheet from 1992 to 2017

The IMBIE team\*

**The Antarctic Ice Sheet is an important indicator of climate change and driver of sea-level rise. Here we combine satellite observations of its changing volume, flow and gravitational attraction with modelling of its surface mass balance to show that it lost  $2,720 \pm 1,390$  billion tonnes of ice between 1992 and 2017, which corresponds to an increase in mean sea level of  $7.6 \pm 3.9$  millimetres (errors are one standard deviation). Over this period, ocean-driven melting has caused rates of ice loss from West Antarctica to increase from  $53 \pm 29$  billion to  $159 \pm 26$  billion tonnes per year; ice-shelf collapse has increased the rate of ice loss from the Antarctic Peninsula from  $7 \pm 13$  billion to  $33 \pm 16$  billion tonnes per year. We find large variations in and among model estimates of surface mass balance and glacial isostatic adjustment for East Antarctica, with its average rate of mass gain over the period 1992–2017 ( $5 \pm 46$  billion tonnes per year) being the least certain.**

The ice sheets of Antarctica hold enough water to raise global sea level by 58 m<sup>1</sup>. They channel ice to the oceans through a network of glaciers and ice streams<sup>2</sup>, each with a substantial inland catchment<sup>3</sup>. Fluctuations in the mass of grounded ice sheets arise owing to differences between net snow accumulation at the surface, meltwater runoff and ice discharge into the ocean. In recent decades, reductions in the thickness<sup>4</sup> and extent<sup>5</sup> of floating ice shelves have disturbed inland ice flow, triggering retreat<sup>6,7</sup>, acceleration<sup>8,9</sup> and drawdown<sup>10,11</sup> of many marine-terminating ice streams. Various techniques have been developed to measure changes in ice-sheet mass, based on satellite observations of their speed<sup>12</sup>, volume<sup>13</sup> and gravitational attraction<sup>14</sup> combined with modelled surface mass balance (SMB)<sup>15</sup> and glacial isostatic adjustment (GIA; the ongoing movement of land associated with changes in ice loading)<sup>16</sup>. Since 1989, there have been more than 150 assessments of ice loss from Antarctica based on these approaches<sup>17</sup>. An inter-comparison of 12 such estimates<sup>18</sup> demonstrated that the three principal satellite techniques provide similar results at the continental scale and, when combined, lead to an estimated mass loss of  $71 \pm 53$  billion tonnes of ice per year (Gt yr<sup>-1</sup>) averaged over the period 1992–2011 (errors are one standard deviation unless stated otherwise). Here, we extend this assessment to include twice as many studies, doubling the overlap period and extending the record to 2017.

## Satellite observations

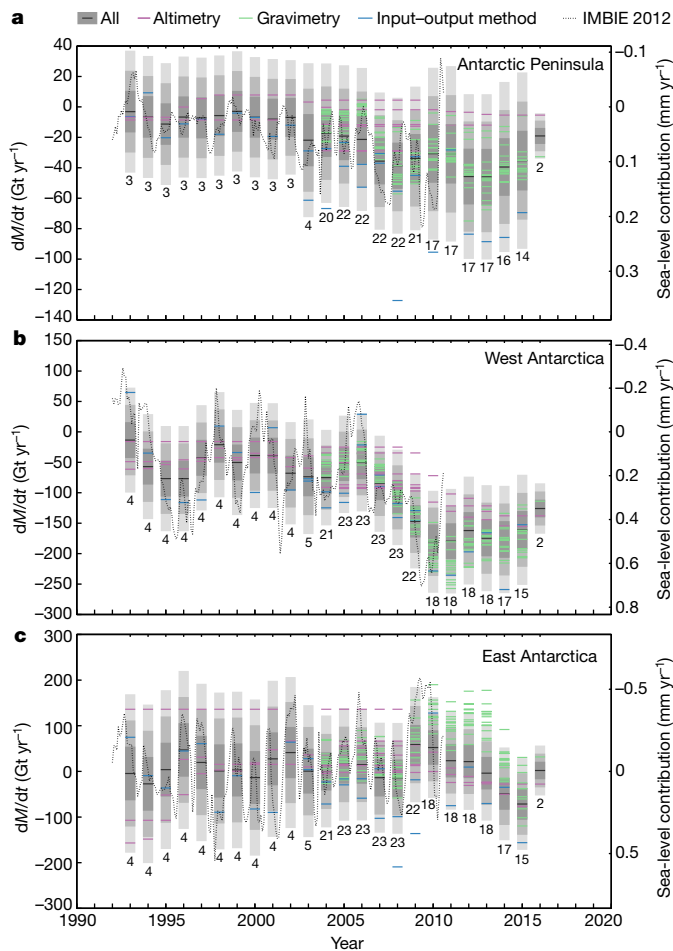
We collated 24 independently derived estimates of ice-sheet mass balance (Fig. 1) that were determined within the period 1992–2017 and based on the techniques of satellite altimetry (seven estimates), gravimetry (15 estimates) or the input–output method (two estimates). Altogether, 24, 24 and 23 individual estimates of mass change were computed within defined geographical limits<sup>3,19</sup> for the East Antarctic Ice Sheet (EAIS), West Antarctic Ice Sheet (WAIS) and Antarctic Peninsula Ice Sheet (APIS), respectively. We compared the rates of ice-sheet mass change (see Methods) over common intervals of time<sup>18</sup>. We then averaged the rates of ice-sheet mass balance using the same class of satellite observations to produce three technique-dependent time series of mass change in each geographical region (see Methods). Within each class, we computed the uncertainty in the annual mass rate as the mean uncertainty of the individual

contributions. The final, reconciled estimate of ice-sheet mass change for each region was computed as the mean of the technique-dependent values available at each epoch (Fig. 1). In computing the associated uncertainty, we assume that the errors for each technique are independent. To estimate the cumulative mass change and its uncertainty (Fig. 2), we integrated the reconciled estimates for each ice sheet and weighted the annual uncertainty by  $1/\sqrt{n}$ , where  $n$  is the number of years since the start of each time series. We computed Antarctic Ice Sheet (AIS) mass trends as the linear sum of the regional trends and the uncertainties in the mass trends as the root-sum-square of the regional uncertainties (Table 1).

## Trends in Antarctic ice-sheet mass

The level of disagreement between individual estimates of ice-sheet mass balance increases with the area of each ice-sheet region, with average per-epoch standard deviations of 11 Gt yr<sup>-1</sup>, 21 Gt yr<sup>-1</sup> and 37 Gt yr<sup>-1</sup> at the APIS, the WAIS and the EAIS, respectively (Fig. 1, Methods). Among the techniques, gravimetric estimates are the most abundant and also the most closely aligned, although their spread increases in East Antarctica, where GIA remains poorly constrained<sup>20</sup> and is least certain when spatially integrated<sup>21–32</sup>, owing to the vast extent of the region. Solutions based on satellite altimetry and the input–output method run for the entire record, roughly twice the duration of the gravimetry time series. Although most (59%) estimates are within one standard deviation of the technique-dependent mean, a few (6%) depart by more than three standard deviations. At the Antarctic Peninsula, the 25-year average rate of ice-sheet mass balance is  $-20 \pm 15$  Gt yr<sup>-1</sup>, with an increase of about 15 Gt yr<sup>-1</sup> in losses since 2000. The strongest signal and trend has occurred in West Antarctica, where rates of mass loss increased from  $53 \pm 29$  Gt yr<sup>-1</sup> to  $159 \pm 26$  Gt yr<sup>-1</sup> between the first and final five years of our survey; the largest increase occurred during the late 2000s when ice discharge from the Amundsen Sea sector accelerated<sup>33</sup>. Both of these regional losses are driven by reductions in the thickness and extent of floating ice shelves, which has triggered the retreat, acceleration and drawdown of marine-terminating glaciers<sup>34</sup>. The least certain result is in East Antarctica, where the average 25-year mass trend is  $5 \pm 46$  Gt yr<sup>-1</sup>. Overall, the AIS lost  $2,720 \pm 1,390$  Gt of ice between 1992 and 2017, an average rate of  $109 \pm 56$  Gt yr<sup>-1</sup>.

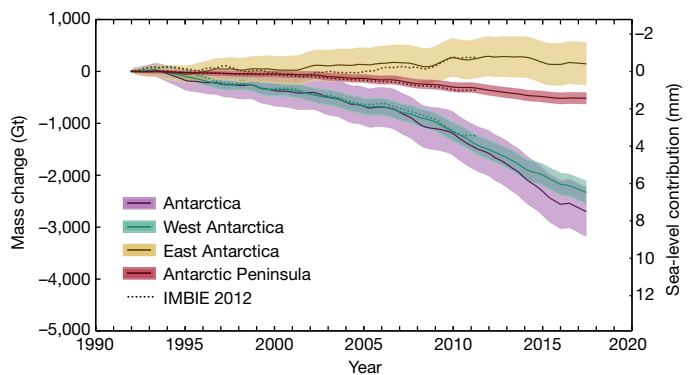
\*A list of authors and their affiliations appears at the end of the paper.



**Fig. 1 | Antarctic Ice Sheet mass balance.** **a–c**, Rate of mass change ( $dM/dt$ ) of the APIS (**a**), WAIS (**b**) and EAIS (**c**), as determined from the various satellite-altimetry (purple), input-output-method (blue) and gravimetry (green) assessments included in this study. In each case,  $dM/dt$  is computed from time series of relative mass change using a three-year window at annual intervals. An average of estimates across each class of measurement technique is also shown for each year (black). The estimated  $1\sigma$ ,  $2\sigma$  and  $3\sigma$  ranges of the class averages are shaded in dark, mid and light grey, respectively; the number of individual mass-balance estimates collated at each epoch is shown below.

## Surface mass balance

Knowledge of the ice-sheet SMB is an essential component of the input–output method, which subtracts solid-ice discharge from net snow accumulation, and aids interpretation of mass trends derived from satellite altimetry and gravimetry. Snowfall is the main driver of temporal and spatial variability in AIS surface mass change<sup>35,36</sup>. Although locally important, spatially integrated sublimation and melt-water runoff are typically one and two orders of magnitude smaller, respectively. In the absence of observation-based maps, AIS SMB is usually taken from atmospheric models, evaluated with in situ and remotely sensed observations<sup>15,37–40</sup>. To assess Antarctic SMB, we compared two global reanalysis products (JRA55 and ERA-Interim) and two regional climate models (RACMO2 and MARv3.6) (see Methods). ERA-Interim is usually regarded as the best-performing reanalysis product over Antarctica, albeit with a dry bias in the interior and overestimated rain fraction<sup>39,41,42</sup>. Spatially averaged accumulation rates peak at the Antarctic Peninsula, and are roughly three and seven times lower in West and East Antarctica, respectively (Extended Data Figs. 2, 3). Compared to the all-model average SMB of  $1,994 \text{ Gt yr}^{-1}$ , the regional climate models give values 4.7% higher and the reanalyses 7% lower. These differences can be attributed to the higher resolution of the regional models, which resolve the steep coastal precipitation



**Fig. 2 | Cumulative Antarctic Ice Sheet mass change.** The cumulative ice-sheet mass changes (solid lines) are determined from the integral of monthly measurement-class averages (for example, the black lines in Fig. 1) for each ice sheet. The estimated  $1\sigma$  uncertainty of the cumulative change is shaded. The dashed lines show the results of a previous assessment<sup>18</sup>.

gradients in greater detail, and to their improved representation of polar processes. The temporal variability of all products is similar and they all agree on the absence of an ice-sheet-wide trend in SMB over the period 1979–2017, which implies that recent mass loss from the AIS is dominated by increased solid-ice discharge into the ocean.

## Glacial isostatic adjustment

Gravimetric estimates of mass change are strongly influenced by the method used to correct for GIA<sup>16</sup>. In this study, six different GIA models were used<sup>21,24,26,30,31,43</sup>. We also assessed nine continent-wide forward-model simulations and two regional model simulations to better understand uncertainties in the GIA signal; we reprocessed the gravimetry estimates of mass balance using the W12a<sup>26</sup> and IJ05\_R2<sup>31</sup> GIA models for comparison with earlier work<sup>18</sup> (see Methods). The net gravitational effect of GIA across Antarctica is positive, and the mean and standard deviation of the continent-wide GIA models ( $54 \pm 18 \text{ Gt yr}^{-1}$ ) is very close to that of the W12a ( $56 \pm 27 \text{ Gt yr}^{-1}$ ) and IJ05\_R2 ( $55 \pm 13 \text{ Gt yr}^{-1}$ ) models. The narrow spread probably reflects the difficulty of quantifying the timing and extent of past ice-sheet change and the absence of lateral variations in Earth rheology within some models<sup>44</sup>. Models predict the greatest rates of solid-Earth uplift ( $5\text{--}7 \text{ mm yr}^{-1}$  on average) in areas where GIA is a substantial component of the regional mass change, such as the Amundsen, Ross and Filchner–Ronne sectors of West Antarctica (see Extended Data Fig. 4), but also the greatest variability (for example, a standard deviation of more than  $10 \text{ mm yr}^{-1}$  in the Amundsen sector). Away from areas with large GIA signals there is low variance among the models and broad agreement with GPS observations<sup>45</sup>. Nevertheless, most models considered here do not account for ice-sheet change during the past few millennia, because it is poorly known. Inaccurate treatment of low-degree harmonics associated with the global GIA signal can also bias gravimetric mass-balance calculations<sup>46</sup>. If the GIA signal includes a transient component associated with recent ice-sheet change, it will bias mass-trend estimates and should be accounted for in future work.

## Outlook

Improvements in assessments of ice-sheet mass balance are still possible. Airborne snow radar<sup>47,48</sup> is a powerful tool for evaluating models of SMB and firn compaction over large spatial (thousands of kilometres) and temporal (centennial) scales, in addition to the ice cores that have traditionally been used<sup>49</sup>. Geological constraints on the ice-sheet history<sup>20</sup> and GPS measurements of contemporary uplift<sup>45,50</sup> enable GIA models to be scrutinized and calibrated. More of both of these types of datasets are needed, especially in East Antarctica. Given their apparent diversity, the spread of models of GIA and SMB should be evaluated in concert with the satellite gravimetry, altimetry and velocity measurements. A reassessment of the satellite measurements acquired during



**Table 1 | Rates of ice-sheet mass change**

	1992–1997 (Gt yr <sup>-1</sup> )	1997–2002 (Gt yr <sup>-1</sup> )	2002–2007 (Gt yr <sup>-1</sup> )	2007–2012 (Gt yr <sup>-1</sup> )	2012–2017 (Gt yr <sup>-1</sup> )	1992–2011 (Gt yr <sup>-1</sup> )	1992–2017 (Gt yr <sup>-1</sup> )
EAIS	11 ± 58	8 ± 56	12 ± 43	23 ± 38	–28 ± 30	13 ± 50	5 ± 46
WAIS	–53 ± 29	–41 ± 28	–65 ± 27	–148 ± 27	–159 ± 26	–73 ± 28	–94 ± 27
APIS	–7 ± 13	–6 ± 13	–20 ± 15	–35 ± 17	–33 ± 16	–16 ± 14	–20 ± 15
AIS	–49 ± 67	–38 ± 64	–73 ± 53	–160 ± 50	–219 ± 43	–76 ± 59	–109 ± 56

Rates were determined from all satellite measurements over various epochs for the EAIS, WAIS and APIS, which combined constitute the AIS. The period 1992–2011 is included for comparison to a previous assessment<sup>18</sup>, which reported mass-balance estimates of 14 ± 43 Gt yr<sup>-1</sup> for the EAIS, –65 ± 26 Gt yr<sup>-1</sup> for the WAIS, –20 ± 14 Gt yr<sup>-1</sup> for the APIS and –71 ± 53 Gt yr<sup>-1</sup> for the AIS. The small differences in our updated estimates for this period are due to our inclusion of more data. Errors are 1 $\sigma$ .

the 1990s would address the imbalance that is present in the current record. Alternative techniques (see, for example, ref.<sup>51</sup>) for combining satellite datasets should be explored, and satellite measurements with common temporal sampling should be contrasted. The ice-sheet mass-balance record should now be separated into the contributions due to short-term fluctuations in SMB and to longer-term trends in glacier ice. In addition to these improvements, continued satellite observations are, of course, essential.

### Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0179-y>

Received: 12 April 2018; Accepted: 24 April 2018;

Published online 13 June 2018.

- Fretwell, P. et al. Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *Cryosphere* **7**, 375–393 (2013).
- Rignot, E., Mouginot, J. & Scheuchl, B. Ice flow of the Antarctic ice sheet. *Science* **333**, 1427–1430 (2011).
- Zwally, H. J., Giovinetto, M. B., Beckley, M. A. & Saba, J. L. Antarctic and Greenland drainage systems. *GSFC Cryospheric Sciences Laboratory* [http://icesat4.gsfc.nasa.gov/cryo\\_data/ant\\_grn\\_drainage\\_systems.php](http://icesat4.gsfc.nasa.gov/cryo_data/ant_grn_drainage_systems.php) (2012).
- Shepherd, A. et al. Recent loss of floating ice and the consequent sea level contribution. *Geophys. Res. Lett.* **37**, L13503 (2010).
- Cook, A. J. & Vaughan, D. G. Overview of areal changes of the ice shelves on the Antarctic Peninsula over the past 50 years. *Cryosphere* **4**, 77–98 (2010).
- Rignot, E., Mouginot, J., Morlighem, M., Seroussi, H. & Scheuchl, B. Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011. *Geophys. Res. Lett.* **41**, 3502–3509 (2014).
- Konrad, H. et al. Net retreat of Antarctic glacier grounding lines. *Nat. Geosci.* **11**, 258–262 (2018).
- Joughin, I., Tulaczyk, S., Bindschadler, R. & Price, S. F. Changes in west Antarctic ice stream velocities: observation and analysis. *J. Geophys. Res. Solid Earth* **107**, 2289 (2002).
- Rignot, E. et al. Accelerated ice discharge from the Antarctic Peninsula following the collapse of Larsen B ice shelf. *Geophys. Res. Lett.* **31**, L18401 (2004).
- Shepherd, A., Wingham, D. J. & Mansley, J. A. D. Inland thinning of the Amundsen Sea sector, West Antarctica. *Geophys. Res. Lett.* **29**, <https://doi.org/10.1029/2001GL014183> (2002).
- Scambos, T. A., Bohlander, J. A., Shuman, C. A. & Skvarca, P. Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. *Geophys. Res. Lett.* **31**, L18402 (2004).
- Rignot, E. & Thomas, R. H. Mass balance of polar ice sheets. *Science* **297**, 1502–1506 (2002).
- Wingham, D. J., Ridout, A. J., Scharroo, R., Arthern, R. J. & Shum, C. K. Antarctic elevation change from 1992 to 1996. *Science* **282**, 456–458 (1998).
- Velicogna, I. & Wahr, J. Measurements of time-variable gravity show mass loss in Antarctica. *Science* **311**, 1754–1756 (2006).
- van Wessem, J. M. et al. Modelling the climate and surface mass balance of polar ice sheets using RACMO2 – part 2: Antarctica (1979–2016). *Cryosphere* **12**, 1479–1498 (2018).
- King, M. A. et al. Lower satellite-gravitymetry estimates of Antarctic sea-level contribution. *Nature* **491**, 586–589 (2012).
- Briggs, K. et al. Charting ice-sheet contributions to global sea-level rise. *Eos* **97**, <https://doi.org/10.1029/2016EO055719> (2016).
- Shepherd, A. et al. A reconciled estimate of ice-sheet mass balance. *Science* **338**, 1183–1189 (2012).
- Rignot, E., Mouginot, J. & Scheuchl, B. Antarctic grounding line mapping from differential satellite radar interferometry. *Geophys. Res. Lett.* **38**, L10504 (2011).
- Bentley, M. J. et al. A community-based geological reconstruction of Antarctic ice sheet deglaciation since the Last Glacial Maximum. *Quat. Sci. Rev.* **100**, 1–9 (2014).
- Peltier, W. R. Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci.* **32**, 111–149 (2004).
- A. G., Wahr, J. & Zhong, S. Computations of the viscoelastic response of a 3-D compressible earth to surface loading: an application to glacial isostatic adjustment in Antarctica and Canada. *Geophys. J. Int.* **192**, 557–572 (2013).
- Sasgen, I. et al. Antarctic ice-mass balance 2003 to 2012: regional reanalysis of GRACE satellite gravimetry measurements with improved estimate of glacial-isostatic adjustment based on GPS uplift rates. *Cryosphere* **7**, 1499–1512 (2013).
- Peltier, W. R., Argus, D. F. & Drummond, R. Space geodesy constrains ice age terminal deglaciation: the global ICE-6G-C (VM5a) model. *J. Geophys. Res. Solid Earth* **120**, 450–487 (2015).
- King, M. A., Whitehouse, P. L. & van der Wal, W. Incomplete separability of Antarctic plate rotation from glacial isostatic adjustment deformation within geodetic observations. *Geophys. J. Int.* **204**, 324–330 (2016).
- Whitehouse, P. L., Bentley, M. J., Milne, G. A., King, M. A. & Thomas, I. D. A new glacial isostatic adjustment model for Antarctica: calibrated and tested using observations of relative sea-level change and present-day uplift rates. *Geophys. J. Int.* **190**, 1464–1482 (2012).
- Spada, G., Melini, D. & Colletti, F. SELEN v2.9.12, <https://geodynamics.org/cgi/software/selen/> (Computational Infrastructure for Geodynamics, 2015).
- Konrad, H., Sasgen, I., Pollard, D. & Klemann, V. Potential of the solid-Earth response for limiting long-term West Antarctic Ice Sheet retreat in a warming climate. *Earth Planet. Sci. Lett.* **432**, 254–264 (2015).
- Briggs, R. D., Pollard, D. & Tarasov, L. A data-constrained large ensemble analysis of Antarctic evolution since the Eemian. *Quat. Sci. Rev.* **103**, 91–115 (2014).
- Ivins, E. R. & James, T. S. Antarctic glacial isostatic adjustment: a new assessment. *Antarct. Sci.* **17**, 541–553 (2005).
- Ivins, E. R. et al. Antarctic contribution to sea level rise observed by GRACE with improved GIA correction. *J. Geophys. Res. Solid Earth* **118**, 3126–3141 (2013).
- Nield, G. A. et al. Rapid bedrock uplift in the Antarctic Peninsula explained by viscoelastic response to recent ice unloading. *Earth Planet. Sci. Lett.* **397**, 32–41 (2014).
- Mouginot, J., Rignot, E. & Scheuchl, B. Sustained increase in ice discharge from the Amundsen Sea Embayment, West Antarctica, from 1973 to 2013. *Geophys. Res. Lett.* **41**, 1576–1584 (2014).
- Shepherd, A., Fricker, H. A. & Farrell, S. L. Trends and connections across the Antarctic cryosphere. *Nature* **558**, <https://doi.org/10.1038/s41586-018-0171-6> (2018).
- Boening, C., Lebsock, M., Landerer, F. & Stephens, G. Snowfall-driven mass change on the East Antarctic ice sheet. *Geophys. Res. Lett.* **39**, L21501 (2012).
- Medley, B. et al. Temperature and snowfall in Western Queen Maud Land increasing faster than climate model projections. *Geophys. Res. Lett.* **45**, 1472–1480 (2018).
- Favier, V. et al. An updated and quality controlled surface mass balance dataset for Antarctica. *Cryosphere* **7**, 583–597 (2013).
- van de Berg, W. J. & Medley, B. Brief Communication: Upper-air relaxation in RACMO2 significantly improves modelled interannual surface mass balance variability in Antarctica. *Cryosphere* **10**, 459–463 (2016).
- Palerm, C. et al. Evaluation of Antarctic snowfall in global meteorological reanalyses. *Atmos. Res.* **190**, 104–112 (2017).
- van Wessem, J. M. et al. Improved representation of East Antarctic surface mass balance in a regional atmospheric climate model. *J. Glaciol.* **60**, 761–770 (2014).
- Bromwich, D. H., Nicolas, J. P. & Monaghan, A. J. An assessment of precipitation changes over Antarctica and the southern ocean since 1989 in contemporary global reanalyses. *J. Clim.* **24**, 4189–4209 (2011).
- Behrangi, A. et al. Status of high-latitude precipitation estimates from observations and reanalyses. *J. Geophys. Res.* **121**, 4468–4486 (2016).
- Klemann, V. & Martinec, Z. Contribution of glacial-isostatic adjustment to the geocentre motion. *Tectonophysics* **511**, 99–108 (2011).
- van der Wal, W., Whitehouse, P. L. & Schrama, E. J. O. Effect of GIA models with 3D composite mantle viscosity on GRACE mass balance estimates for Antarctica. *Earth Planet. Sci. Lett.* **414**, 134–143 (2015).
- Martín-Español, A. et al. An assessment of forward and inverse GIA solutions for Antarctica. *J. Geophys. Res. Solid Earth* **121**, 6947–6965 (2016).
- Caron, L. et al. GIA model statistics for GRACE hydrology, cryosphere, and ocean science. *Geophys. Res. Lett.* **45**, 2203–2212 (2018).
- Medley, B. et al. Constraining the recent mass balance of Pine Island and Thwaites glaciers, West Antarctica, with airborne observations of snow accumulation. *Cryosphere* **8**, 1375–1392 (2014).

48. Lewis, G. et al. Regional Greenland accumulation variability from Operation IceBridge airborne accumulation radar. *Cryosphere* **11**, 773–788 (2017).
49. Thomas, E. R. et al. Regional Antarctic snow accumulation over the past 1000 years. *Clim. Past* **13**, 1491–1513 (2017).
50. Thomas, I. D. et al. Widespread low rates of Antarctic glacial isostatic adjustment revealed by GPS observations. *Geophys. Res. Lett.* **38**, L22302 (2011).
51. Wahr, J., Wingham, D. & Bentley, C. A method of combining ICESat and GRACE satellite data to constrain Antarctic mass balance. *J. Geophys. Res. Solid Earth* **105**, 16279–16294 (2000).

**Acknowledgements** This work is an outcome of the ESA–NASA Ice Sheet Mass Balance Inter-comparison Exercise. A.S. was additionally supported by a Royal Society Wolfson Research Merit Award and by the ESA Climate Change Initiative.

**Reviewer information** *Nature* thanks R. Bell and C. Hulbe for their contribution to the peer review of this work.

**Author contributions** A.S. and E.I. designed and led the study. E.R., B.S., M.v.d.B., I.V. and P.W. led the input–output-method, altimetry, SMB, gravimetry and GIA experiments, respectively. G.M. and M.E.P. performed the data collation and analysis. A.S., E.I., K.B., G.K., M.H., I.J., H.K., M.M., J.M., S.N., I.O., M.E.P., T.P., E.R., I.S., T.Sc., N.S., T.S.I., B.S., I.V., M.v.W. and P.W. wrote and edited the manuscript. All authors participated in the data interpretation and commented on the manuscript.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0179-y>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0179-y>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### The IMBIE team

Andrew Shepherd<sup>1\*</sup>, Erik Ivins<sup>2</sup>, Eric Rignot<sup>3</sup>, Ben Smith<sup>4</sup>, Michiel van den Broeke<sup>5</sup>, Isabella Velicogna<sup>3</sup>, Pippa Whitehouse<sup>6</sup>, Kate Briggs<sup>1</sup>, Ian Joughin<sup>4</sup>, Gerhard Krinner<sup>7</sup>, Sophie Nowicki<sup>8</sup>, Tony Payne<sup>9</sup>, Ted Scambos<sup>10</sup>, Nicole Schlegel<sup>2</sup>, Geruo A<sup>3</sup>, Cécile Agosta<sup>11</sup>, Andreas Ahlstrøm<sup>12</sup>, Greg Babonis<sup>13</sup>, Valentina Barletta<sup>14</sup>, Alejandro Blazquez<sup>15</sup>, Jennifer Bonin<sup>16</sup>, Beata Csatho<sup>13</sup>, Richard Cullather<sup>17</sup>, Denis Felikson<sup>18</sup>, Xavier Fettweis<sup>11</sup>, Rene Forsberg<sup>14</sup>, Hubert Gallee<sup>7</sup>, Alex Gardner<sup>2</sup>, Lin Gilbert<sup>19</sup>, Andreas Groh<sup>20</sup>, Brian Gunter<sup>21</sup>, Edward Hanna<sup>22</sup>, Christopher Harig<sup>23</sup>, Veit Helm<sup>24</sup>, Alexander Horvath<sup>25</sup>, Martin Horwath<sup>20</sup>, Shfaqat Khan<sup>14</sup>, Kristian K. Kjeldsen<sup>12,26</sup>, Hannes Konrad<sup>1</sup>, Peter Langen<sup>27</sup>, Benoit Lecavalier<sup>28</sup>, Bryant Loomis<sup>8</sup>, Scott Luthcke<sup>8</sup>, Malcolm McMillan<sup>1</sup>, Daniele Melini<sup>29</sup>, Sebastian Mernild<sup>30,31,32</sup>, Yara Mohajerani<sup>3</sup>,

Philip Moore<sup>33</sup>, Jeremie Mouginot<sup>3,7</sup>, Gorka Moyano<sup>34</sup>, Alan Muir<sup>19</sup>, Thomas Nagler<sup>35</sup>, Grace Nield<sup>6</sup>, Johan Nilsson<sup>2</sup>, Brice Noel<sup>5</sup>, Ines Otsaka<sup>1</sup>, Mark E. Pattle<sup>34</sup>, W. Richard Peltier<sup>36</sup>, Nadege Pie<sup>18</sup>, Roelof Rietbroek<sup>37</sup>, Helmut Rott<sup>35</sup>, Louise Sandberg-Sørensen<sup>14</sup>, Ingo Sasgen<sup>24</sup>, Himanshu Save<sup>18</sup>, Bernd Scheuch<sup>3</sup>, Ernst Schrama<sup>38</sup>, Ludwig Schröder<sup>20</sup>, Ki-Weon Seo<sup>39</sup>, Sebastian Simonsen<sup>14</sup>, Tom Slater<sup>1</sup>, Giorgio Spada<sup>40</sup>, Tyler Sutterley<sup>3</sup>, Matthieu Talpe<sup>41</sup>, Lev Tarasov<sup>28</sup>, Willem Jan van de Berg<sup>5</sup>, Wouter van der Wal<sup>38</sup>, Melchior van Wessem<sup>5</sup>, Bramha Dutt Vishwakarma<sup>42</sup>, David Wiese<sup>2</sup> & Bert Wouters<sup>5</sup>

<sup>1</sup>Centre for Polar Observation and Modelling, University of Leeds, Leeds, UK. <sup>2</sup>NASA Jet Propulsion Laboratory, Pasadena, CA, USA. <sup>3</sup>Department of Earth System Science, University of California, Irvine, CA, USA. <sup>4</sup>Department of Earth and Space Sciences, University of Washington, Seattle, WA, USA. <sup>5</sup>Institute for Marine and Atmospheric Research, Utrecht University, Utrecht, The Netherlands. <sup>6</sup>Department of Geography, Durham University, Durham, UK. <sup>7</sup>Institute of Environmental Geosciences, Université Grenoble Alpes, Grenoble, France. <sup>8</sup>Cryospheric Sciences Laboratory, NASA Goddard Space Flight Centre, Greenbelt, MD, USA. <sup>9</sup>School of Geographical Sciences, University of Bristol, Bristol, UK. <sup>10</sup>National Snow and Ice Data Centre, University of Colorado, Boulder, CO, USA. <sup>11</sup>Department of Geography, University of Liège, Liège, Belgium. <sup>12</sup>Geological Survey of Denmark and Greenland, Copenhagen, Denmark. <sup>13</sup>Department of Geology, State University of New York at Buffalo, Buffalo, NY, USA. <sup>14</sup>DTU Space, National Space Institute, Technical University of Denmark, Kongens Lyngby, Denmark. <sup>15</sup>Spatial Geophysics and Oceanography Studies Laboratory, Toulouse, France. <sup>16</sup>College of Marine Sciences, University of South Florida, Tampa, FL, USA. <sup>17</sup>Earth System Science Interdisciplinary Centre, NASA Goddard Space Flight Centre, Greenbelt, MD, USA. <sup>18</sup>Centre for Space Research, University of Texas, Austin, TX, USA. <sup>19</sup>Mullard Space Science Laboratory, University College London, Holmbury St Mary, UK. <sup>20</sup>Institute for Planetary Geodesy, Technische Universität Dresden, Dresden, Germany. <sup>21</sup>Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. <sup>22</sup>School of Geography, University of Lincoln, Lincoln, UK. <sup>23</sup>Department of Geosciences, University of Arizona, Tucson, AZ, USA. <sup>24</sup>Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany. <sup>25</sup>Institute of Astronomical and Physical Geodesy, Technical University Munich, Munich, Germany. <sup>26</sup>Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark. <sup>27</sup>Danish Meteorological Institute, Copenhagen, Denmark. <sup>28</sup>Department of Physics and Physical Oceanography, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada. <sup>29</sup>Section of Seismology and Tectonophysics, National Institute of Geophysics and Volcanology, Rome, Italy. <sup>30</sup>Nansen Environmental and Remote Sensing Centre, Bergen, Norway. <sup>31</sup>Faculty of Engineering and Science, Western Norway University of Applied Sciences, Sogndal, Norway. <sup>32</sup>Direction of Antarctic and Sub-Antarctic Programs, Universidad de Magallanes, Punta Arenas, Chile. <sup>33</sup>School of Civil Engineering and Geosciences, Newcastle University, Newcastle upon Tyne, UK. <sup>34</sup>isardSAT, Barcelona, Spain. <sup>35</sup>ENVEO, Innsbruck, Austria. <sup>36</sup>Department of Physics, University of Toronto, Toronto, Ontario, Canada. <sup>37</sup>Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany. <sup>38</sup>Department of Space Engineering, Delft University of Technology, Delft, The Netherlands. <sup>39</sup>Department of Earth Science Education, Seoul National University, Seoul, South Korea. <sup>40</sup>Institute of Physics, University of Urbino “Carlo Bo”, Urbino, Italy. <sup>41</sup>Aerospace Engineering Sciences, Centre for Astrodynamics Research, University of Colorado, Boulder, CO, USA. <sup>42</sup>Geodetic Institute, University of Stuttgart, Stuttgart, Germany. \*e-mail: a.shepherd@leeds.ac.uk



## METHODS

**Data.** No statistical methods were used to predetermine sample size.

We analyse five groups of data: mass-balance estimates determined from satellite altimetry, gravimetry and the input–output method, and model estimates of SMB and GIA. We compute the datasets using common spatial and temporal domains to facilitate their aggregation, according to previously reported methods (see Supplementary Table 1). In total, 24 mass-balance datasets were included. The data include 25 years of satellite-radar-altimeter measurements, 24 years of satellite input–output-method measurements and 14 years of satellite-gravimetry measurements (Extended Data Fig. 1). Among these data are estimates of ice-sheet mass balance for each ice sheet derived from each satellite technique. In comparison to the first IMBIE assessment<sup>18</sup>, new satellite missions, updated methodologies and improvements in geophysical corrections have contributed to an increase in the quantity, duration and overlap period of data used here. In addition, two new experiment groups have assessed 11 GIA and 4 SMB models. The complete list of datasets is provided in Supplementary Table 1.

**Drainage basins.** In this assessment, we analyse mass trends using two sets of ice-sheet drainage basins (Extended Data Fig. 2) to ensure consistency with those used in the first IMBIE assessment<sup>18</sup> and to evaluate an updated definition tailored towards assessments using the input–output method. The first drainage-basin set was delineated using surface elevation maps derived from ICESat-1 on the basis of the provenance of the ice and includes 27 basins<sup>5</sup>. The second set was updated to consider other factors such as the direction of ice flow and includes 18 basins in Antarctica<sup>2,19</sup>. To assess the effect of the different sets on the estimates of ice-sheet mass balance, we compared mass-balance determinations between the two delineations of ice-sheet drainage basins. This evaluation was facilitated by seven estimates (altimetry or gravimetry) determined using both sets. At the scale of the major ice-sheet divisions, the delineations produce similar total extents. By far the largest differences occur in the delineation (or definition) of East and West Antarctica, owing to differences in the position of the ice divide that separate them. Within these regions, the root-mean-square difference between 26 pairs of estimates of ice-sheet mass balance computed using the two drainage-basin sets is  $8.7 \text{ Gt yr}^{-1}$ . This difference is small in comparison to the certainty of individual assessments of ice-sheet mass balance.

**Computing rates of mass change.** The raw satellite mass-balance data are time series of either relative mass change  $\Delta M(t)$  or the rate of mass change  $dM(t)/dt$ , plus their associated uncertainty, integrated over at least one of the ice-sheet regions defined in the standard drainage-basin sets. In the case of  $\Delta M(t)$ , the time series represents the change in mass through time relative to some nominal reference value. The duration and sampling frequency of the time series was not restricted. In practice, few mass time series were of  $\Delta M(t)$  and  $dM(t)/dt$ . Because the inter-comparison exercise is based on comparing and aggregating  $dM(t)/dt$ , a common solution was implemented to derive  $dM(t)/dt$  values from datasets of  $\Delta M(t)$  only. Each  $\Delta M(t)$  time series was used to generate a time-varying estimate of  $dM(t)/dt$ ,  $d[\Delta M(t)]/dt = dM(t)/dt$ , and an estimate of the associated uncertainty, using a consistent approach. Time-varying estimates of  $dM(t)/dt$  were computed by applying a sliding fixed-period window to the  $\Delta M(t)$  time series. At each node, defined by the sampling period of the input time series,  $dM(t)/dt$  and its standard error  $\sigma_{dM(t)/dt}$  were estimated by fitting a linear trend to data within the window using a weighted least-squares approach, with each point weighted by its respective error variance  $\sigma_{\Delta M(t)}^2$ . The regression error  $\sigma_{dM(t)/dt}$  incorporates measurement errors and model structural error due to any variability that deviates from linear trends in ice mass, and may be a conservative estimate in locations where such deviation is present. Time series of  $dM(t)/dt$  computed using this approach were truncated by half the moving-average window period. When integrated, the  $dM(t)/dt$  time series correspond to a low-pass-filtered version of the original  $\Delta M(t)$  time series. This linear regression assumes that uncertainties are uncorrelated; however, the smoothing that we apply during the calculation of the trend causes data points to be correlated during several epochs beyond the sliding window.

**Surface mass balance.** Ice-sheet SMB comprises various processes governed by the interaction of the superficial snow and firn layers with the atmosphere. A direct mass exchange occurs via precipitation and surface sublimation. Snow drift and the formation of meltwater and its subsequent refreezing or retention redistribute mass spatially or lead to further mass loss via erosion and sublimation or via runoff. Here we compare a range of SMB products. Four SMB model solutions were considered for Antarctica (Extended Data Table 1): two regional models (RACMO2.3<sup>40</sup> and MARv3.6<sup>52</sup>) and two global reanalysis products (JRA55<sup>53</sup> and ERA-Interim<sup>54</sup>). The two regional climate models agree well in terms of their spatially integrated SMB, apart from the Peninsula where there is an offset of about  $10 \text{ Gt per month}$  between them (Extended Data Fig. 3). However, the reanalysis products underestimate the average SMB compared to the regional climate models by  $200\text{--}350 \text{ Gt yr}^{-1}$ . Our SMB assessment illustrates that products of similar class (climate models or reanalysis products) agree well, suggesting that groupings of their output may be appropriate. However, we found that model resolution is important when

estimating SMB and its components, because contributions that differed by only the spatial resolution yielded differences at the regional scale.

**Glacial isostatic adjustment.** GIA is the delayed response of the solid Earth to changes in time-variable surface loading through the growth and decay of ice sheets, and associated changes in sea level. Because GIA contributes to changes in the ice-sheet surface elevation and gravity field, it must be accounted for in measurements of the change in elevation and gravity for the purpose of isolating the contribution solely caused by ice-sheet imbalance. Here we compare different solutions derived from continuum-mechanical forward modelling to inform the interpretation of the satellite altimetry and gravimetry data that depend on the correction and to advise future assessments. Twelve GIA contributions were received that cover Antarctica (Extended Data Table 2), ten of which are global models<sup>22–29,31</sup> and two of which are regional models<sup>32</sup>. Because a broad array of data may be used to constrain GIA forward models, we anticipate spread in the predictions.

Here we assess the degree of similarity between the various GIA model solutions. We identified areas of enhanced present-day vertical surface motion and (dis-)agreement between contributions by averaging the uplift rates over the contributions and computing the respective standard deviations (Extended Data Fig. 4). In some cases, it was necessary to estimate the GIA contribution to gravimetric mass trends; this was done using common geographical masks and truncation and a standardized treatment of low-degree harmonics. In Antarctica, the Amundsen Sea sector and the regions covered by the Ross and Filchner Ronne ice shelves stand out as having both high uplift rates ( $5\text{--}7 \text{ mm yr}^{-1}$  on average) and high variability in uplift rates (peaking at more than  $10 \text{ mm yr}^{-1}$  standard deviation in the Amundsen sector) among the models considered. Elsewhere in coastal regions, uplift occurs at more moderate rates (about  $2 \text{ mm yr}^{-1}$  on average); the interior of East Antarctica exhibits slow subsidence. In these regions, the average signal is accompanied by relatively low variance among the GIA models ( $0\text{--}1.5 \text{ mm yr}^{-1}$  standard deviation). None of the models fully captures portions of the uplift that are observed to be very large (see, for example, ref. <sup>55</sup>); hence, we anticipate a bias towards low values for the GIA correction averaged over such regions. In areas of low mantle viscosity, however, such as part of the WAIS, the GIA signal related to the Last Glacial Maximum may be overpredicted, and it is not clear whether a bias exists at the continental scale.

Differences between the model predictions arise for various additional reasons. Technical differences in the modelling approach, for example, relating to the consideration of self-gravitation, ocean loading, rotational feedback and compressibility, are most important at the global scale, but may explain only small differences among the regional models. Differing treatment of ice and ocean loading in regions that have experienced marine-based grounding-line retreat during the last glacial cycle may explain the differences in model predictions for the ICE\_6G\_C/VM5a combination (see Supplementary Table 1). Some small differences should be expected when comparing models that use spherical-harmonic and finite-element approaches. Looking beyond consideration of the model physics, larger differences arise owing to the various approaches used to determine the two principal unknowns associated with forward modelling of GIA—ice history and Earth rheology. There is no generally accepted ‘best approach’ to determining these inputs, and useful advances can be made by comparing the results of complementary approaches. In the models considered here, approaches to determining the ice history include dynamical ice-sheet modelling, coupled ice-sheet–GIA modelling, tuning to fit geodetic constraints, tuning to fit geological constraints and use of direct observations of historical ice-sheet change. When defining the rheological properties of the solid Earth, most studies have opted to use a Maxwell rheology to define a radially symmetric Earth; however, the use of a power-law rheology or a fully three-dimensional Earth model to capture the spatial complexity of mantle properties is increasingly popular. An intermediate approach used in many of the datasets included here has been to develop a regional GIA model that reflects local Earth structure. Such models can be tuned, albeit imperfectly, to provide as accurate a representation of GIA in that region as is possible. However, it remains a difficult and important challenge to incorporate these regional studies into a global framework. Finally, although four of the GIA models that we consider provide a measure of uncertainty, and several studies have used an ensemble modelling approach<sup>23,29</sup>, an important future goal for the GIA modelling community is the inclusion of robust error estimates for all model predictions.

To compare the GIA models, we used Stokes coefficients that relate to their gravitational signal to determine the approximate magnitude of the effect of applying each correction to GRACE data (Extended Data Table 2). This is a preliminary assessment, because the effect of applying a GIA correction depends also on the methods used to process the GRACE data. Moreover, an agreement on the modelling of feedbacks has so far not been reached within the GIA community, leading to a large spread in the modelled degree-2 coefficients and possibly a strong bias when a correction is applied that is inconsistent with the GRACE observations (up to about  $40 \text{ Gt yr}^{-1}$ ). In addition, none of the current GIA datasets includes estimates

of the GIA-induced geocentre motion (degree-1 coefficients). Therefore, we omit degree-1 and -2 coefficients in our assessment of the GIA-induced apparent mass change. From models that represent GIA in Antarctic only, we estimate that this omission could change the apparent mass-change value by up to 20%; however, this potential error is not currently included in the GIA error budget. There is relatively good agreement between the ten models that cover all of Antarctica (Extended Data Table 2); the estimated GIA contribution ranges from  $+12 \text{ Gt yr}^{-1}$  to  $+81 \text{ Gt yr}^{-1}$ ; the mean value is  $56 \text{ Gt yr}^{-1}$ . Although the solution from ref. <sup>44</sup> is a notable outlier, it is the only one to account for three-dimensional variations in Earth's rheology. It will be interesting to compare this result with other such models that are under development.

Two of the GIA models<sup>32,56</sup> are regional: although they cannot be compared with the continental-scale models directly, the magnitude of their signals is nonetheless included for interest.

**Mass-balance intra-comparison.** First, we compare estimates of mass change within each of the three geodetic-technique experiment groups to assess the degree to which results from common techniques concur and to derive individual, aggregated estimates of mass change from each technique. In each case, we compare estimated rates of mass change derived from a common technique over a common geographical region and over the full period of the respective datasets. Where datasets were computed using both drainage-basin definitions, we present the arithmetic mean of the two estimates. This is justified because the choice of drainage-basin set has a very small (less than  $10 \text{ Gt yr}^{-1}$ ) effect on estimates of mass balance at the ice-sheet scale and even less at the regional scale. Within each experiment group, we perform an unweighted average of all individual data to obtain a single estimate of the rate of mass change per ice sheet for each geodetic technique. In a few cases, it was not possible to determine time-varying rates of mass change from individual estimates, because only constant rates of mass change and constant cumulative mass changes were supplied. Although the effect of averaging these datasets with time-varying solutions is to dampen the temporal variability present within the series of finer resolution, they are retained for completeness. We estimate the uncertainty of the average mass trends that emerge from each experiment group as the average of the errors associated with each individual estimate at each epoch.

To aid comparison, we (i) compute time-variable rates of mass change and their associated uncertainty over successive 36-month periods stepped in one-month intervals from time-varying cumulative mass changes, and (ii) average rates of mass change over one-year periods to remove signals associated with seasonal cycles. Time-varying rates of mass change are truncated at the start and end of each series to reflect the half-width of the time interval over which rates are computed, although this period is recovered on integration to cumulative mass changes. The extent to which we are able to analyse differences in mass-balance solutions that emerge from common satellite approaches is limited by the mismatch in temporal resolution of the individual datasets, which makes methodological and sampling differences difficult to separate.

**Gravimetry mass-balance intra-comparison.** Within the gravimetry experiment group, we assessed 15 estimates of mass balance derived from the GRACE satellites, in entirety spanning the period July 2002 to September 2016. Of these datasets, four<sup>57–60</sup> were derived with direct imposition of the GRACE level-1 K-band range data. These impositions result in four different, independently derived, mascon approaches. Other methods often refer to 'mascon analysis', but are conducted on post-spherical-harmonic expansions and without imposing the level-1 K-band range data. We distinguish the later methods, referring to them as 'post-spherical-harmonic mascons'. Eleven contributions are derived from monthly spherical-harmonic solutions of the global gravity field using different approaches<sup>55,56,61–66</sup>, which can be loosely classified as (i) region-integration approaches<sup>55,65,66</sup>, (ii) post-spherical-harmonic mascon approaches<sup>56,61–63</sup>, (iii) forward-modelling approaches<sup>62,64</sup>, which essentially involve modelling of mass change with iterative comparison to the GRACE-derived signal, and (iv) approaches that use Slepian functions<sup>67</sup>. One final estimate<sup>68</sup> made use of satellite altimetry data; although this estimate was excluded from our gravity ensemble average because it is a hybrid solution, it is presented alongside the gravimetry-only results for comparison. No restrictions were imposed on the choice of GIA correction; among the GRACE solutions we consider six different models<sup>21,24,26,30,31,43</sup>. However, we did assess a wider set of nine continent-wide forward models and two regional models to better understand uncertainties in the GIA signal.

In total, there were 15 estimates of mass balance for each of the APIS, WAIS and EAIS. All were time-varying, cumulative mass-change solutions—the primary GRACE observable—and we computed time-varying rates of mass change from these data. Combining all of the individual mass-balance estimates, the effective (average) temporal resolution of the aggregated solution is one year. Further details of the gravimetry datasets and methods are included in Supplementary Table 1.

In Extended Data Fig. 5 we show a comparison of the rates of mass change obtained from all gravimetry mass-balance solutions, calculated over the three

main ice-sheet regions. At individual epochs, differences between time-varying rates of mass change are generally less than  $50 \text{ Gt yr}^{-1}$  in each ice-sheet region, and typically in the range  $10\text{--}20 \text{ Gt yr}^{-1}$ . Over the full period of the data, individual rates of mass balance for the APIS, WAIS and EAIS vary between  $-80 \text{ Gt yr}^{-1}$  and  $+10 \text{ Gt yr}^{-1}$ ,  $-260 \text{ Gt yr}^{-1}$  and  $-20 \text{ Gt yr}^{-1}$ , and  $-120 \text{ Gt yr}^{-1}$  and  $+200 \text{ Gt yr}^{-1}$ , respectively. Considering all of the gravimetry data (Extended Data Table 3), the standard deviation of mass trends estimated during the period 2005–2015 is less than  $24 \text{ Gt yr}^{-1}$  in all three ice-sheet regions, with the largest spread occurring in the EAIS. In all three ice-sheet regions, the spread of individual mass balance estimates is well represented by the mean, considering the uncertainties of the individual and aggregated datasets.

**Altimetry mass-balance intra-comparison.** We assessed seven radar- and laser- altimetry-derived AIS mass-balance datasets, in entirety spanning the period April 1992 to July 2017. In total, six estimates of mass change were for the APIS, seven were for the EAIS and seven were for the WAIS. Of these, four included data from radar altimetry and six from laser altimetry. Various techniques were used to derive the elevation and mass trends<sup>69–75</sup>. Only two of the altimetry datasets were time series of cumulative mass change, from which we computed time-varying rates of mass change. The remaining altimetry datasets were constant rates of mass change, which appear in our altimetry average as time-invariant solutions. The period over which altimetry rates of mass change were computed ranged from 2 years to 24 years. In consequence, the aggregated dataset has a temporal resolution that is lower than annual. Including all individual mass-balance datasets, the effective (average) temporal resolution of the aggregated solution is 3.3 years. Further details of the altimetry datasets and methods are included in Supplementary Table 1.

With a few exceptions, rates of mass change determined from radar and laser altimetry tend to differ by less than  $100 \text{ Gt yr}^{-1}$  at all times in each ice-sheet region (Extended Data Fig. 5). The main exceptions are in the EAIS, where one estimate<sup>74</sup> reports mass trends that are roughly  $100 \text{ Gt yr}^{-1}$  more positive than all others during the ERS and ICESat periods, and in the WAIS, where two estimates<sup>71,74</sup> report rates that are about  $70 \text{ Gt yr}^{-1}$  less negative than the others during the ICESat period. Among the remaining datasets, the closest agreement occurs at the APIS, where mass trends agree to within  $30 \text{ Gt yr}^{-1}$  at all times; the poorest agreement occurs at the EAIS, where mass trends depart by up to  $100 \text{ Gt yr}^{-1}$ . The largest differences are between datasets that are constant in time during periods where rapid changes in mass balance occur in the annually resolved time series, suggesting that a proportion of the difference is due to their poor temporal resolution. Mass-balance solutions from the relatively short (six-year) ICESat mission also appear to show larger spreads compared to those determined from longer (decade-scale) radar- altimetry missions. This larger spread is due in part to differences in the bias-correction models applied to ICESat data<sup>74,76–78</sup> and in part to the large influence of firn densification on altimetry measurements over short periods, which have been corrected for using different models. Firn-densification models are generally not applied to mass-balance solutions determined from radar altimetry. Further analysis of the corrections for bias between ICESat campaigns and firn compaction is required to establish the statistical significance of the differences and to reduce their collective uncertainty. Comparing rates of mass change (Extended Data Table 3), the average standard deviation of all mass trends at each epoch over the common period 2005–2015 is less than  $54 \text{ Gt yr}^{-1}$  in all four ice-sheet regions. The largest spread between the individual values occurs in the EAIS. Other than this sector, the individual estimates lie close to the ensemble average, considering the respective uncertainty of the measurements.

**Input-output-method intra-comparison.** Although the input–output method is the most direct measure of changes in mass fluxes, a key difficulty is that, to assess mass balance, it must differentiate between two large numbers—one for annual SMB and the other for discharge plus grounding-line migrations—and deal appropriately with the error budgets of both. A consequence of this complexity is that few input–output-method datasets exist at the ice-sheet scale. Here we collate just two input–output-method datasets, both based on the same method<sup>79</sup>—far fewer than were considered for altimetry and gravimetry. The first dataset spans the period 1992–2010<sup>18</sup>; the second is limited to the period 2002–2016. The same SMB model (RACMO2.3) was used in both assessments. Further details of the input–output-method datasets and methods are included in Supplementary Table 1.

We compare the two datasets during the period 2002–2010 (when the datasets overlap; Extended Data Table 3). The smallest differences (up to  $30 \text{ Gt yr}^{-1}$ ) arise in the APIS and the WAIS; the largest differences (up to  $70 \text{ Gt yr}^{-1}$ ) occur at the EAIS. In all cases, the average difference between estimates of mass balance derived from each dataset is comparable to the estimated uncertainty. Including both datasets, rates of mass balance over the period 1992–2016 for the APIS, WAIS and EAIS range from  $-125 \text{ Gt yr}^{-1}$  to  $+25 \text{ Gt yr}^{-1}$ ,  $-300 \text{ Gt yr}^{-1}$  to  $+100 \text{ Gt yr}^{-1}$  and  $-200 \text{ Gt yr}^{-1}$  to  $+200 \text{ Gt yr}^{-1}$ , respectively (Extended Data Fig. 5). The origin of the differences between the two datasets requires further investigation.

**Ice-sheet mass-balance inter-comparison.** To assess the degree to which the satellite techniques concur, we used the aggregated time series from each



geodetic-technique experiment group to compute changes in ice-sheet mass balance within common geographical regions and over a common interval of time (the overlap period). We calculate the aggregated time series as the arithmetic mean of all available rates of ice-sheet mass balance derived from the same satellite technique at each available epoch. We used the individual ice sheets and their integrals as common geographical regions. The maximum duration of the overlap period is limited to the 14-year interval (2002–2016) when all three satellite techniques were optimally operational. However, we also considered the availability of mass-balance datasets, which leads us to select the period 2003–2010 as the optimal interval (see Fig. 1). When the aggregated mass-balance data from all three experiment groups are degraded to a common temporal resolution of 36 months, the time series are on average well correlated ( $0.5 < r^2 < 0.9$ ) at the APIS and WAIS. At the EAIS, however, the aggregated altimetry mass-balance time series are poorly correlated ( $r^2 < 0.1$ ) in time with the aggregated gravimetry and input–output-method data. Possible explanations for this include the relatively high short-term variability in mass fluctuations in this region, the relatively low trend in mass and the heterogeneous temporal resolution of the aggregated altimetry dataset. Over longer periods, marked increases in the rate of mass loss from the WAIS are also recorded in all three satellite datasets.

Because the comparison period is long in relation to the timescales over which SMB fluctuations typically occur, their potential effect on the overall inter-comparison is reduced. The closest agreement between individual estimates of ice-sheet mass balance occurs at the APIS and the WAIS, where the standard deviation across all techniques is  $15\text{--}41\text{ Gt yr}^{-1}$  (Extended Data Table 4). The greatest departure occurs at the EAIS, where the input–output-method and gravimetry estimates of mass balance differ by about  $80\text{ Gt yr}^{-1}$  and the standard deviation of all three estimates is about  $40\text{ Gt yr}^{-1}$ . This high degree of variance is expected because of the relatively large size of the region, the small amplitude of signals and the poor independent controls on coastal SMB. When compared to the inter-technique mean and standard deviation, all estimates of ice-sheet mass balance determined from the individual satellite techniques are now in agreement, given their respective uncertainties. In contrast to the first IMBIE assessment<sup>18</sup>, this finding also now holds at continental and global scales. We therefore conclude that estimates of mass balance determined from independent geodetic techniques agree when compared to their respective uncertainties.

Several noteworthy patterns in the distribution of mass-balance estimates determined during the overlap period (2003–2010) merit further discussion. Estimates of mass balance derived from satellite altimetry and gravimetry agree to within  $15\text{ Gt yr}^{-1}$  on average and with the mean of all three techniques, in all ice-sheet regions. By contrast, estimates of mass balance determined from the input–output method are  $55\text{ Gt yr}^{-1}$  more negative on average than the mean in all ice-sheet regions. However, despite the bias, the input–output-method estimates remain in agreement because their estimated uncertainties are relatively large (approximately three times larger than those of the other techniques). A more detailed analysis of the primary and ancillary datasets is required to establish whether this bias is significant or systematic.

**Ice-sheet mass-balance integration.** We combined estimates of ice-sheet mass balance derived from each geodetic-technique experiment group to produce a single, reconciled estimate, following the same approach as for the first assessment. This estimate was computed as the arithmetic mean of the average rates of mass change from each experiment group, within the regions of interest and at the time periods for which the experiment-group mass trends were determined. We estimated the uncertainty of the mass-balance data using the following approach. Within each experiment group, we estimated the uncertainty of mass trends as the average of the errors associated with each individual estimate and the uncertainty of reconciled rates of mass change (see, for example, Table 1) as the root-mean-square of the uncertainties associated with mass trends from each experiment group. When summing mass trends of multiple ice sheets, the combined uncertainty was estimated as the root-sum-square of the uncertainties for each region. Finally, to estimate the cumulative uncertainty of mass changes over time, we weighted the annual uncertainty by  $1/\sqrt{n}$ , where  $n$  is the number of years since the start of each time series, and summed the weighted annual uncertainties over time<sup>80</sup>.

Across the full 25-year survey, the average rate of mass balance of the AIS was  $-109 \pm 56\text{ Gt yr}^{-1}$  (Table 1). To investigate inter-annual variability, we also calculated mass trends during successive five-year intervals. Whereas the APIS and WAIS each lost mass throughout the entire survey period, the EAIS experienced alternating periods of mass loss and mass gain, probably driven by inter-annual fluctuations in SMB. The rate of mass loss from the WAIS has increased over time owing to accelerated ice discharge in the Amundsen Sea sector<sup>33,47,73,81–83</sup>. The largest increase—a doubling of the rate of ice loss—occurred between the periods 2002–2007 and 2007–2012 (Table 1). Overall, the WAIS accounts for the vast majority of ice-mass losses from Antarctica. At the APIS, rates of ice-mass loss since the early 2000s are notably higher than during the previous decade,

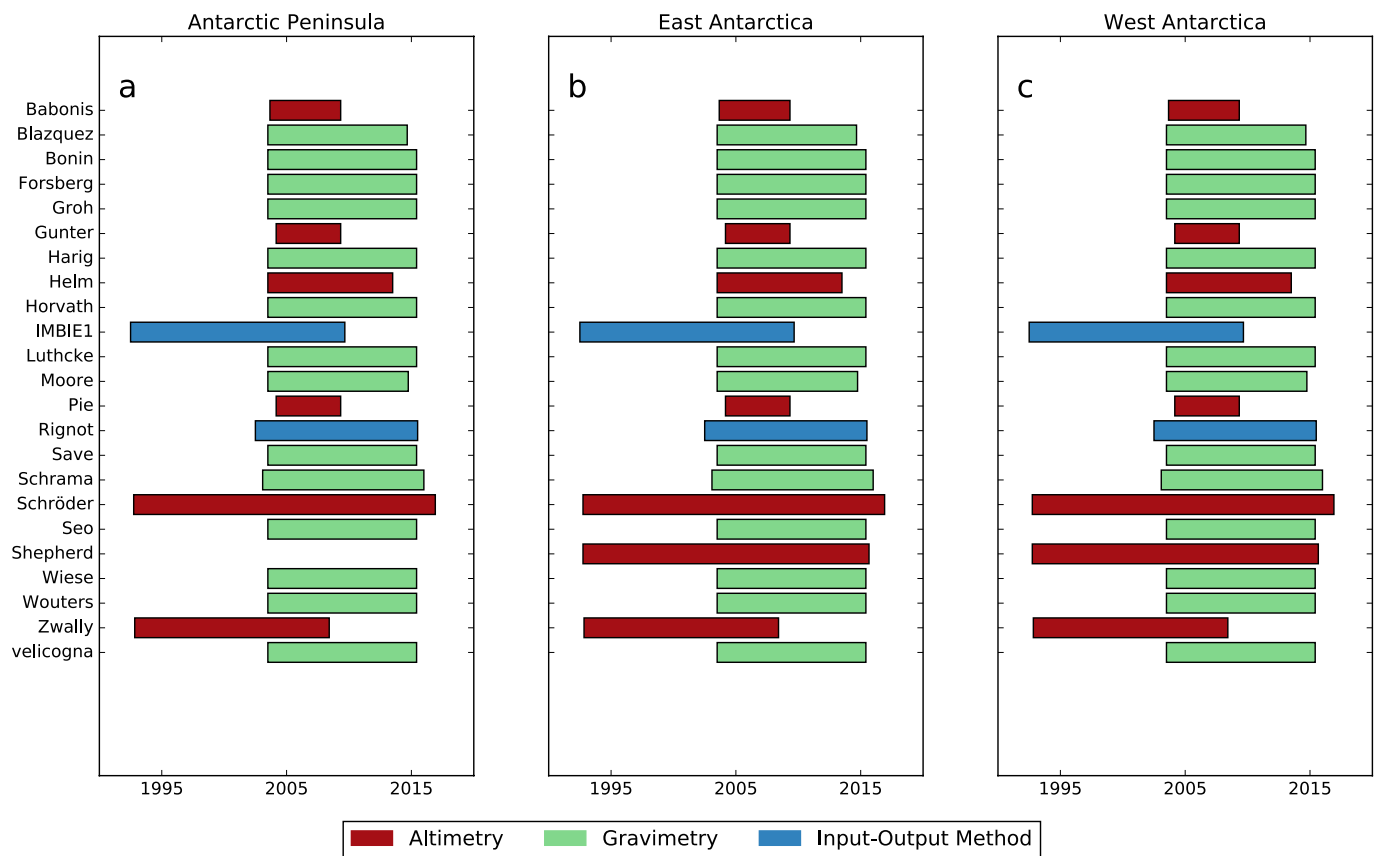
consistent with observations of surface lowering<sup>71,73</sup> and increased ice flow in southerly glacier catchments<sup>84</sup>. The approximate state of balance of the wider EAIS suggests that the reported dynamic thinning of the Totten and Cook glaciers<sup>85,86</sup> has been offset by accumulation gains elsewhere<sup>87</sup>.

**Data availability.** The final mass-balance datasets generated in this study are freely available at <http://www.imbie.org/data-downloads>.

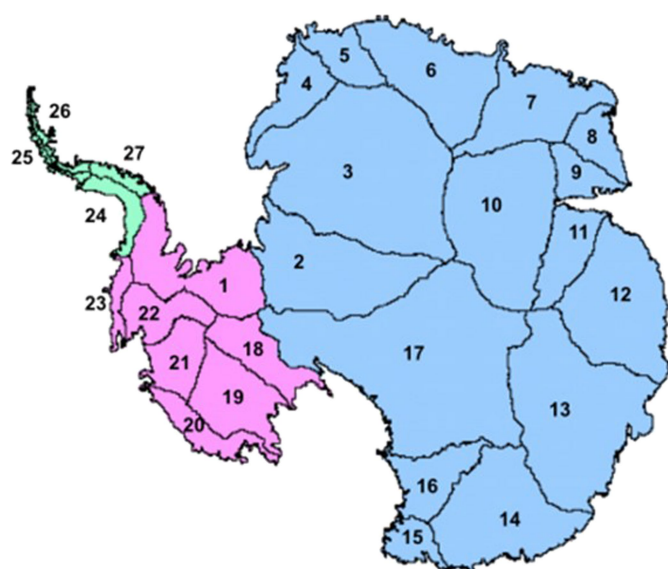
52. Fettweis, X. et al. Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR. *Cryosphere* **7**, 469–489 (2013).
53. Kobayashi, S. et al. The JRA-55 reanalysis: general specifications and basic characteristics. *J. Meteorol. Soc. Jpn.* **93**, 5–48 (2015).
54. Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
55. Groh, A. & Horwath, M. The method of tailored sensitivity kernels for GRACE mass change estimates. *Geophys. Res. Abstr.* **18**, 12065 (2016).
56. Barletta, V. R., Sørensen, L. S. & Forsberg, R. Scatter of mass changes estimates at basin scale for Greenland and Antarctica. *Cryosphere* **7**, 1411–1432 (2013).
57. Luthcke, S. B. et al. Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *J. Glaciol.* **59**, 613–631 (2013).
58. Andrews, S. B., Moore, P. & King, M. A. Mass change from GRACE: a simulated comparison of Level-1B analysis techniques. *Geophys. J. Int.* **200**, 503–518 (2015).
59. Save, H., Bettadpur, S. & Tapley, B. D. High-resolution CSR GRACE RL05 mascons. *J. Geophys. Res. Solid Earth* **121**, 7547–7569 (2016).
60. Watkins, M. M., Wiese, D. N., Yuan, D. N., Boening, C. & Landerer, F. W. Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *J. Geophys. Res. Solid Earth* **120**, 2648–2671 (2015).
61. Schrama, E. J. O., Wouters, B. & Rietbroek, R. A mascon approach to assess ice sheet and glacier mass balances and their uncertainties from GRACE data. *J. Geophys. Res. Solid Earth* **119**, 6048–6066 (2014).
62. Seo, K. W. et al. Surface mass balance contributions to acceleration of Antarctic ice mass loss during 2003–2013. *J. Geophys. Res. Solid Earth* **120**, 3617–3627 (2015).
63. Velicogna, I., Sutterley, T. C. & van den Broeke, M. R. Regional acceleration in ice mass loss from Greenland and Antarctica using GRACE time-variable gravity data. *Geophys. Res. Lett.* **41**, 8130–8137 (2014).
64. Wouters, B., Bamber, J. L., van den Broeke, M. R., Lenaerts, J. T. M. & Sasgen, I. Limits in detecting acceleration of ice sheet mass loss due to climate variability. *Nat. Geosci.* **6**, 613–616 (2013).
65. Blazquez, A. et al. Exploring the uncertainty in GRACE estimates of the mass redistributions at the Earth surface. *Implications for the global water and sea level budgets*. (submitted).
66. Horwath, A. G. *Retrieving Geophysical Signals from Current and Future Satellite Missions*. PhD thesis, Tech. Univ. Munich (2017).
67. Harig, C. & Simons, F. J. Mapping Greenland's mass loss in space and time. *Proc. Natl Acad. Sci. USA* **109**, 19934–19937 (2012).
68. Rietbroek, R., Brunnabend, S. E., Kusche, J. & Schröder, J. Resolving sea level contributions by identifying fingerprints in time-variable gravity and altimetry. *J. Geodyn.* **59–60**, 72–81 (2012).
69. Babonis, G. S., Csatho, B. & Schenk, T. Mass balance changes and ice dynamics of Greenland and Antarctic ice sheets from laser altimetry. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* Vol. XLII-B8 (eds Zdimar, V. et al.) 481–487 (ISPRS, 2016).
70. Felikson, D. et al. Comparison of elevation change detection methods from ICESat altimetry over the Greenland Ice Sheet. *IEEE Trans. Geosci. Remote Sens.* **55**, 5494–5505 (2017).
71. Helm, V., Humbert, A. & Miller, H. Elevation and elevation change of Greenland and Antarctica derived from CryoSat-2. *Cryosphere* **8**, 1539–1559 (2014).
72. Ewert, H. et al. Precise analysis of ICESat altimetry data and assessment of the hydrostatic equilibrium for subglacial Lake Vostok, East Antarctica. *Geophys. J. Int.* **191**, 557–568 (2012).
73. McMillan, M. et al. Increased ice losses from Antarctica detected by CryoSat-2. *Geophys. Res. Lett.* **41**, 3899–3905 (2014).
74. Zwally, H. J. et al. Mass gains of the Antarctic ice sheet exceed losses. *J. Glaciol.* **61**, 1019–1036 (2015).
75. Gunter, B. C. et al. Empirical estimation of present-day Antarctic glacial isostatic adjustment and ice mass change. *Cryosphere* **8**, 743–760 (2014).
76. Scambos, T. & Shuman, C. Comment on 'Mass gains of the Antarctic ice sheet exceed losses' by H. J. Zwally and others. *J. Glaciol.* **62**, 599–603 (2016).
77. Zwally, H. J. et al. Response to Comment by T. SCAMBOS and C. SHUMAN (2016) on 'Mass gains of the Antarctic ice sheet exceed losses' by H. J. Zwally and others (2015). *J. Glaciol.* **62**, 990–992 (2016).
78. Richter, A. et al. Height changes over subglacial Lake Vostok, East Antarctica: insights from GNSS observations. *J. Geophys. Res. Earth Surf.* **119**, 2460–2480 (2014).
79. Rignot, E., Velicogna, I., van den Broeke, M. R., Monaghan, A. & Lenaerts, J. Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophys. Res. Lett.* **38**, L05503 (2011).
80. Stocker, T. F. et al. (eds) *Climate Change 2013: the Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* Ch. 4 (Cambridge Univ. Press, New York, 2013).
81. Boman, J. et al. Antarctic outlet glacier mass change resolved at basin scale from satellite gravity gradiometry. *Geophys. Res. Lett.* **41**, 5919–5926 (2014).

82. Konrad, H. et al. Uneven onset and pace of ice-dynamical imbalance in the Amundsen Sea embayment, West Antarctica. *Geophys. Res. Lett.* **44**, 910–918 (2017).
83. Gardner, A. S. et al. Increased West Antarctic and unchanged East Antarctic ice discharge over the last 7 years. *Cryosphere* **12**, 521–547 (2018).
84. Hogg, A. E. et al. Increased ice flow in Western Palmer Land linked to ocean melting. *Geophys. Res. Lett.* **44**, 4159–4167 (2017).
85. Pritchard, H. D., Arthern, R. J., Vaughan, D. G. & Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461**, 971–975 (2009).
86. Li, X., Rignot, E., Morlighem, M., Mouginot, J. & Scheuchl, B. Grounding line retreat of Totten Glacier, East Antarctica, 1996 to 2013. *Geophys. Res. Lett.* **42**, 8049–8056 (2015).
87. Lenaerts, J. T. M. et al. Recent snowfall anomalies in Dronning Maud Land, East Antarctica, in a historical and future climate perspective. *Geophys. Res. Lett.* **40**, 2684–2688 (2013).
88. Pollard, D. & Deconto, R. M. Description of a hybrid ice sheet-shelf model, and application to Antarctica. *Geosci. Model Dev.* **5**, 1273–1295 (2012).
89. Martinec, Z. Spectral-finite element approach to three-dimensional viscoelastic relaxation in a spherical earth. *Geophys. J. Int.* **142**, 117–141 (2000).

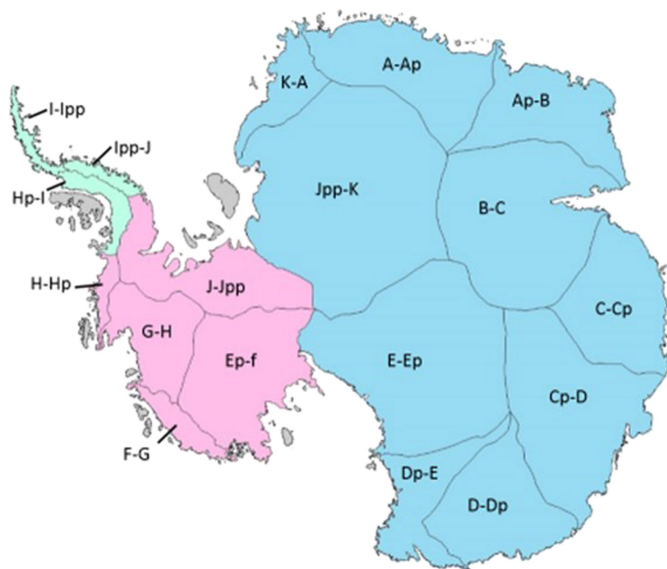




**Extended Data Fig. 1 | Datasets of ice-sheet mass balance included in our assessment.** Details about the datasets are provided in Supplementary Table 1. Some datasets did not encompass all three ice sheets.

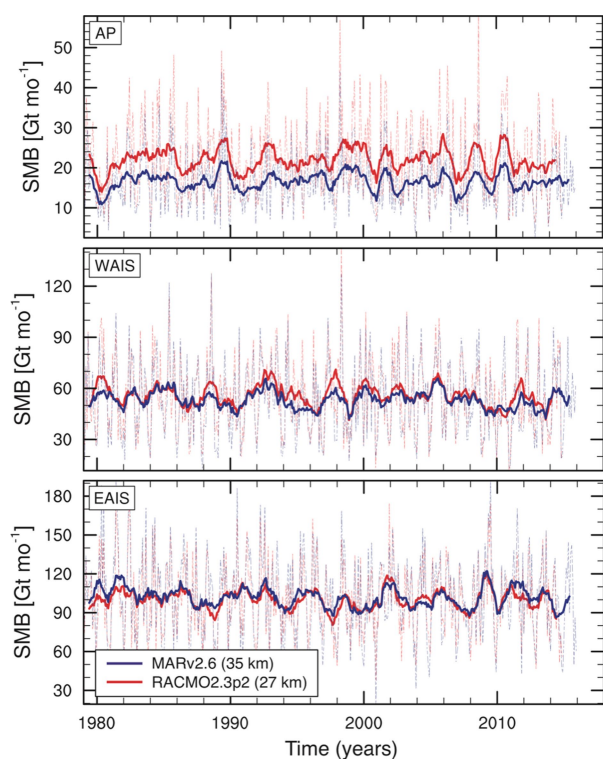


**Extended Data Fig. 2 | Ice-sheet drainage basins.** AIS drainage basins are determined according to the definitions of ref. <sup>3</sup> (left) and refs <sup>2,19</sup> (right). Basins that fall within the Antarctic Peninsula, West Antarctica and East Antarctica are shown in green, pink and blue, respectively. For the definition from ref. <sup>3</sup>, the Antarctic Peninsula, West Antarctica and

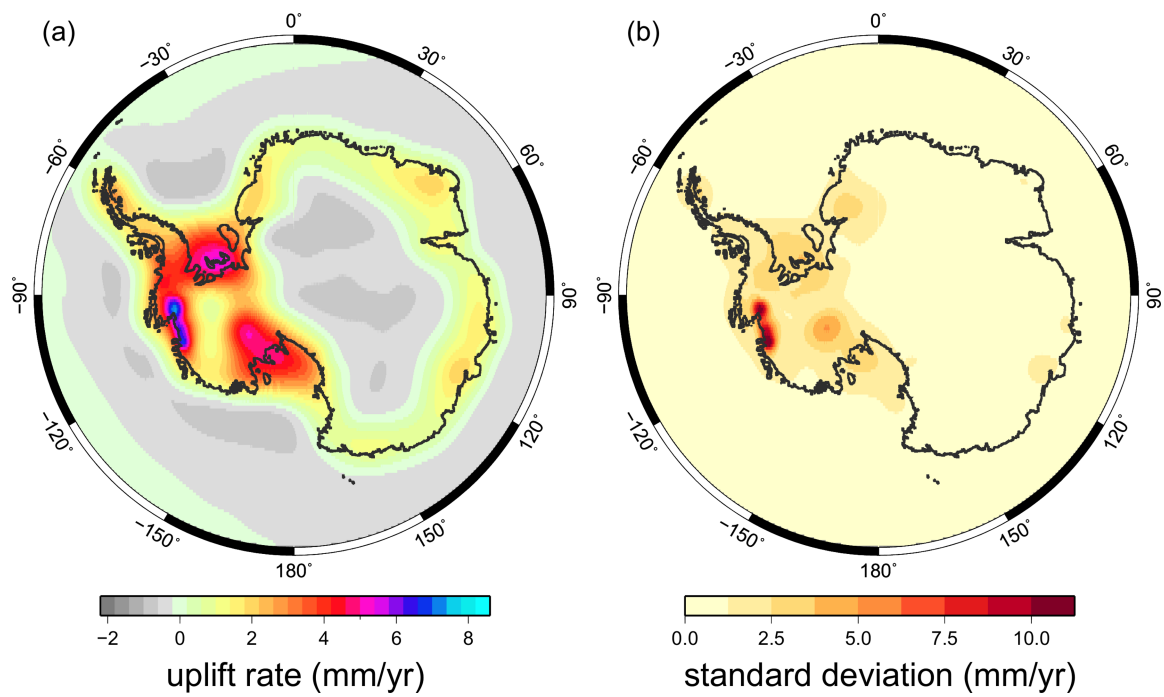


East Antarctica basins cover areas of 227,725 km<sup>2</sup>, 1,748,200 km<sup>2</sup> and 9,909,800 km<sup>2</sup>, respectively. For the definition from refs <sup>2,19</sup>, the Antarctic Peninsula, West Antarctica and East Antarctica basins cover areas of 232,950 km<sup>2</sup>, 2,039,525 km<sup>2</sup> and 9,620,225 km<sup>2</sup>, respectively.

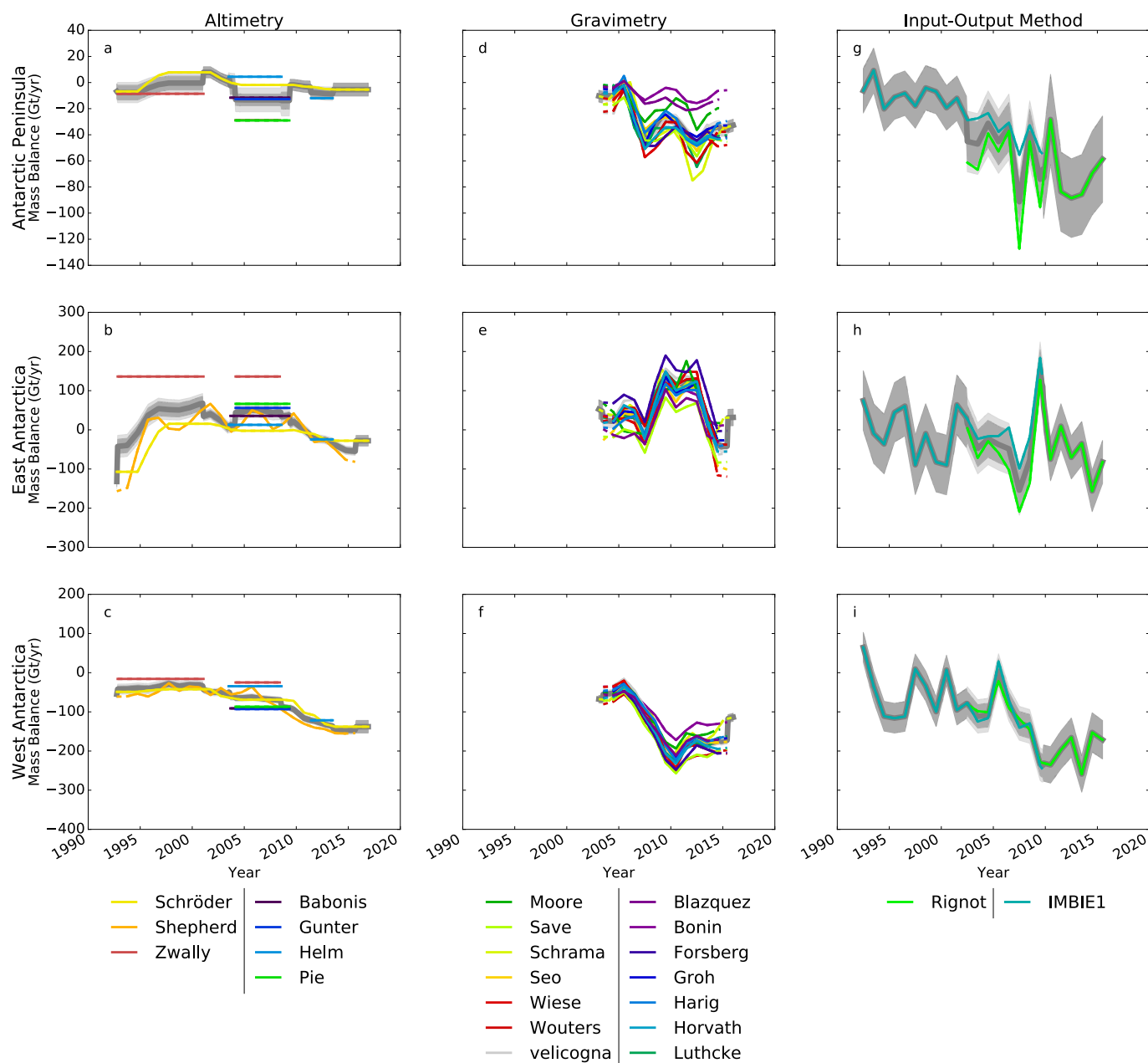




**Extended Data Fig. 3 | Temporal variations in AIS SMB.** We show time series of integrated SMB in AIS drainage regions<sup>2,19</sup> from the MARv2.6 (blue) and RACMO2.3p2 (red) models. Solid lines are annual averages of the monthly data (dashed lines). mo, month.



**Extended Data Fig. 4 | Modelled GIA beneath the AIS.** **a**, Bedrock uplift rates in Antarctica averaged over the GIA model solutions used in this assessment. **b**, The corresponding standard deviations.



**Extended Data Fig. 5 | Individual rates of ice-sheet mass balance.** a–i, Mass-balance estimates were determined from satellite altimetry (a–c), gravimetry (d–e) and the input–output method (g–i) for the Antarctic Peninsula (a, d, g), East Antarctica (b, e, h) and West

Antarctica (c, f, i). The light-grey shading shows the estimated  $1\sigma$  uncertainty relative to the ensemble average. The standard error of the mean solutions, per epoch, is shown in mid-grey.



Extended Data Table 1 | Spatially averaged AIS SMB

Model	Class	Area (10 <sup>6</sup> km <sup>2</sup> )	Grid	SMB (Gt/yr)
RACMO2.3	RCM	12.30	27km	2004
RACMO2.3p2	RCM	12.30	27km	2107
MARv3.6.40	RCM	12.32	35km	2150
ERA-Interim	GCM	12.20	80km	1900
JRA55	GCM	12.24	55km	1807

Estimates of the average SMB over the period 1980–2012 were derived from regional climate models (RCM) and global reanalyses (GCM). Data were evaluated using the drainage basins from refs <sup>2,19</sup>.

Extended Data Table 2 | GIA model details

Model	Publication <sup>a</sup>	Ice sheet	Earth model <sup>b</sup>	Ice model <sup>c</sup>	GIA model <sup>d</sup>	Constraint data <sup>e</sup>	Mass change (Gt/yr)
A13	<sup>8</sup>	AIS	VM5a (1D) <sup>i</sup>	ICE-6G_C	SH, C, RF, SG, OL	As for ICE-6G_C	+68 <sup>‡</sup>
AGE1a	<sup>9</sup>	AIS	ensemble of regional 1D models	Own model: ice thickness scaled to fit GPS	SH(256), UQ	GPS	+48 ± 14 <sup>‡</sup>
DIEM/ANT1D.0	<sup>10</sup>	AIS	1D (90,0.5,20)	Dynamically coupled model <sup>j</sup>	SH(170) <sup>k</sup> ; dynamically coupled model	GPS, RSL	+49 <sup>‡</sup>
ICE-6G_C (VM5a)	<sup>5</sup>	AIS	VM5a (1D) <sup>i</sup>	ICE-6G_C	SH(1024)	GPS, RSL, Earth rotation	+72 <sup>‡</sup>
ICE-6G_D (VM5a)	<sup>5</sup>	AIS	VM5a (1D) <sup>i</sup>	ICE-6G_D <sup>f</sup>	SH(512)	GPS, RSL, Earth rotation	+62 <sup>‡</sup>
SL-dry-4mm/W12	<sup>11</sup>	AIS	3D, power-law rheology	Combination of W12 and ICE-5G	FE, IC, xRF	GPS, RSL, seismic velocities (earth model)	+12 <sup>‡</sup>
W12a	<sup>12</sup>	AIS	1D (120,1,10)	Own model: dynamic, time slice	SH(256), C, RF, SG, OL, UQ	GPS, RSL, ice extent & thickness	+56 ± 27 <sup>‡</sup>
SELEN 4	<sup>13</sup>	AIS	VM5a (3-layer average of 1D model) <sup>i</sup>	ICE-6G_C	SELEN4: SH(128), IC, RF, SG, OL	As for ICE-6G_C	+81 <sup>‡</sup>
GLAC1-D	<sup>14</sup>	AIS	VM5a (1D) <sup>i</sup>	Own coupled model: dynamic, from ensemble	SH(512), IC, SG, OL, xRF	RSL, ice extent & thickness, present ice sheet	+55 <sup>‡</sup>
IJ05_R2	<sup>15</sup>	AIS	1D (65,0.2,4)	Own model	SH(256), IC, SG, OL, UQ	GPS, ice extent & thickness	+55 ± 13 <sup>‡</sup>
NAP_N14	<sup>16</sup>	nAPIS <sup>g</sup>	1D (130, 0.0007,0.4,10)	Own model: 1995-present	SH(1195), C, SG, xRF, xOL	GPS, altimetry & DEM difference (ice model)	+3 <sup>‡</sup>
ASE14G (L60S186)	<sup>16</sup>	AS <sup>h</sup>	1D (60, 0.00398,0.0158,0.025)	Own model: 1900-present	SH(1195), C, SG	GPS, altimetry (ice model)	+19 <sup>‡</sup>

<sup>‡</sup>Regional changes in mass associated with the GIA signal were determined from the model data.

<sup>‡</sup>Regional changes in mass associated with the GIA signal were calculated as an indicative rate using spherical-harmonic degrees 3 to 90.

<sup>a</sup>Main publication<sup>5,8–16</sup> listed; supporting publications are provided in Supplementary Table 1.

<sup>b</sup>Model from main publication unless otherwise stated. Comma-separated values refer to properties of a radially varying (1D, one-dimensional) Earth model: the first value is lithosphere thickness (km); other values reflect mantle viscosity ( $\times 10^{21}$  Pa s) for specific layers; see relevant publications for details.

<sup>c</sup>Ice model covers at least the Last Glacial Maximum to present, unless otherwise indicated.

<sup>d</sup>GIA model details: SH, spherical harmonic (maximum degree indicated in parentheses); FE, finite element; C, compressible; IC, incompressible; RF, rotational feedback; SG, self-gravitation; OL, ocean loading; x, feature not included; UQ, uncertainty quantified.

<sup>e</sup>RSL, relative sea-level data; GPS rates were all corrected for the elastic response to contemporary ice mass change.

<sup>f</sup>Different to ICE-6G\_C in Antarctica, owing to the use of BEDMAP2<sup>1</sup> topography in that region.

<sup>g</sup>Model relates to GIA in the northern Antarctic Peninsula (nAPIS) only.

<sup>h</sup>Model relates to GIA in the Amundsen Sea (AS) embayment only.

<sup>i</sup>Earth model from ref. <sup>24</sup>.

<sup>j</sup>Ice model from ref. <sup>88</sup>.

<sup>k</sup>GIA model from ref. <sup>89</sup>.

Extended Data Table 3 | Features of mass-balance datasets included in our assessment

Region	Technique	Span (years)	Temporal resolution (months)	dM/dt range (Gt/yr)	dM/dt error (Gt/yr)	dM/dt standard deviation (Gt/yr)
APIS	Gravimetry	2005 to 2015	1 to 12	-39 to -9	1 to 24	10
WAIS	Gravimetry	2005 to 2015	1 to 12	-177 to -114	1 to 30	16
EAIS	Gravimetry	2005 to 2015	1 to 12	+11 to +107	2 to 35	24
APIS	Altimetry	1992 to 2017	1 to 8.25	-29 to -3	2 to 17	12
WAIS	Altimetry	1992 to 2017	1 to 8.25	-97 to -25	4 to 39	27
EAIS	Altimetry	1992 to 2017	1 to 8.25	-11 to +136	10 to 52	54
APIS	Mass Budget	1992 to 2016	1	-120 to +20	30	35
WAIS	Mass Budget	1992 to 2016	1	-250 to +100	2	61
EAIS	Mass Budget	1992 to 2016	1	-200 to +200	65	95

Details shown include the maximum span, temporal sampling, amplitude, estimated error and standard deviation at each epoch.



**Extended Data Table 4 | Aggregated estimates of ice-sheet mass balance from satellite altimetry, gravimetry and the input–output method**

<b>Region</b>	<b>Altimetry mass balance (Gt/yr)</b>	<b>Gravimetry mass balance (Gt/yr)</b>	<b>Mass budget mass balance (Gt/yr)</b>	<b>Average mass balance (Gt/yr)</b>
EAIS	$37 \pm 18$	$47 \pm 18$	$-35 \pm 65$	$15 \pm 41$
WAIS	$-70 \pm 8$	$-101 \pm 9$	$-115 \pm 43$	$-93 \pm 26$
APIS	$-10 \pm 9$	$-23 \pm 5$	$-51 \pm 24$	$-27 \pm 15$
AIS	$-43 \pm 21$	$-76 \pm 20$	$-201 \pm 82$	$-105 \pm 51$

In this comparison, the data were averaged over the period 2003–2010. The arithmetic mean of each individual result is also shown for the given regions, along with the combined imbalance of the AIS, calculated as the sum of estimates from the constituent regions.

# Trends and connections across the Antarctic cryosphere

Andrew Shepherd<sup>1\*</sup>, Helen Amanda Fricker<sup>2</sup> & Sinead Louise Farrell<sup>3</sup>

**Satellite observations have transformed our understanding of the Antarctic cryosphere. The continent holds the vast majority of Earth's fresh water, and blankets swathes of the Southern Hemisphere in ice. Reductions in the thickness and extent of floating ice shelves have disturbed inland ice, triggering retreat, acceleration and drawdown of marine-terminating glaciers. The waxing and waning of Antarctic sea ice is one of Earth's greatest seasonal habitat changes, and although the maximum extent of the sea ice has increased modestly since the 1970s, inter-annual variability is high, and there is evidence of longer-term decline in its extent.**

At the height of austral winter, Antarctica and the surrounding ocean are covered in a 31.6 million km<sup>2</sup> cap of ice (Fig. 1). Of this, approximately 18.5 million km<sup>2</sup> is formed as sea ice when the ocean freezes<sup>1</sup>, 11.9 million km<sup>2</sup> is a near-permanent ice sheet resting on land or the sea floor<sup>2</sup> and 1.6 million km<sup>2</sup> is contained within long-lived ice shelves that are floating extensions of the continental ice<sup>3</sup>. All of Antarctica's ice is mobile, driven by gravity and, where it is afloat, by the atmosphere and the ocean (Fig. 1). Each element plays a unique role in the climate system; for example, the grounded ice is Earth's primary freshwater reservoir<sup>4</sup>, the ice shelves are a major source of ocean fresh water<sup>5</sup>, and the sea ice is an important factor in the planetary albedo<sup>6</sup>.

The greatest fluctuation in the extent of ice cover in the Southern Hemisphere is due to the seasonal cycle of sea ice formation, which is less than a metre thick on average<sup>7</sup>, and reduces to one-sixth of its peak area in summer<sup>8</sup>. The decadal trend in Antarctic sea ice extent has, nevertheless, been modest<sup>9</sup>, and the most striking contemporary changes have occurred in other elements of the regional cryosphere. For example, the grounded ice sheet is estimated to have lost  $2,720 \pm 1390$  Gt of its mass between 1992 and 2017<sup>10</sup>, and its peripheral ice shelves are thinning in numerous sectors<sup>3,11–13</sup> and collapsing<sup>14,15</sup> at the Antarctic Peninsula. These trends reflect global and regional environmental forcing and are related through a variety of processes, each of which is now better understood thanks to the array of satellite observations that has been acquired over recent decades.

Here we analyse the satellite record to examine continental- and regional-scale trends in the Antarctic cryosphere, including fluctuations in the extent, thickness and movement of sea ice, ice shelves and the grounded ice sheet. We show that spaceborne measurements have allowed key events in Earth's recent climate history to be charted in remarkable detail—including the collapse of ice shelves at the Antarctic Peninsula and the drawdown of glacier ice from West Antarctica—and have illuminated the key processes that are driving contemporary change.

## Grounded ice

Fluctuations in the mass of the Antarctic ice sheet arise from differences between the net snow accumulation and ice discharge. In recent decades, a variety of techniques have been developed to measure changes in the speed, elevation and weight of the grounded ice. Airborne radar measurements show that the Antarctic ice sheet is up to 4,897 m thick, and has the potential to raise global sea level by 58 m were it to be

rapidly discharged<sup>4</sup>. It overlies terrain of variable geology and relief, and this has influenced both its formation and its contemporary dynamics. The continental-scale pattern of ice flow was first inferred from cartographic<sup>16</sup> and, more recently, satellite altimeter<sup>17</sup> records of the ice sheet surface elevation. On this basis, it has been determined that most of Antarctica's ice is routed into the Southern Ocean through around 30 glaciers and ice streams (Fig. 1), each draining a substantial inland catchment<sup>2</sup>.

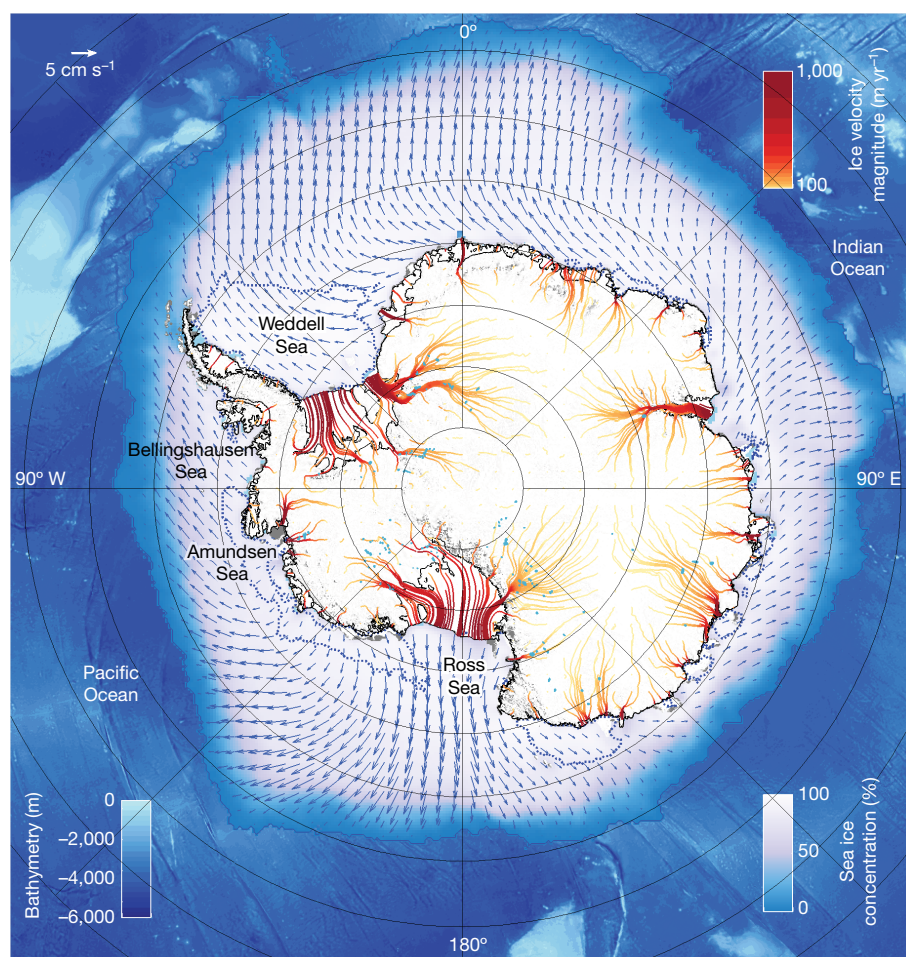
## Grounded ice imbalance

The stability of Antarctica's ice can be assessed by tracking the movement of its principal glaciers and ice streams. Although such glaciers and ice streams are few in number, they are vast, putting this task beyond the scope of ground surveys. The first remote measurements of ice motion were made possible by repeat satellite optical imagery<sup>18</sup> and, subsequently, by synthetic aperture radar interferometry<sup>19</sup>. Thanks to step increases in the quantity of satellite image acquisitions over time, systematic surveys of ice flow across and around the continent have now been completed<sup>20</sup>, revealing anomalous behaviour in much of Marie Byrd Land<sup>21,22</sup> and also at isolated sites at the Siple Coast<sup>23</sup>, at the Antarctic Peninsula<sup>24–26</sup>, and in East Antarctica<sup>27,28</sup>. In most of these places, the pace of ice flow has increased during the satellite era, and, when considered as a whole, the rate of ice discharge from Antarctica exceeds inland snow accumulation<sup>29</sup>.

In addition to changes in ice discharge, fluctuations in ice sheet mass can be detected through satellite measurements of the volume<sup>30–32</sup> and gravitational attraction<sup>33,34</sup> of the ice sheets. Although all three methods lead to similar results at the continental scale, each approach has its merits, and they are now viewed as being complementary. So far, there have been over 150 individual assessments of ice loss from Antarctica based on these approaches<sup>35</sup> and, when collated<sup>10,36</sup>, these studies show that the continent has contributed  $7.6 \pm 3.9$  mm to global sea levels since 1992. Two-fifths ( $3.0 \pm 0.6$  mm) of this loss occurred during the past five years<sup>10</sup>. Although the rate of ice loss from the entire Antarctic ice sheet has changed little during the satellite record, speedup of glacier flow in the Amundsen Sea sector has led to accelerated losses from this region<sup>37,38</sup>.

Satellite radar altimetry is an especially powerful tool for ice sheet glaciology, because the technique can be used to resolve the detailed pattern of imbalance across individual glacier catchments (for example, ref. <sup>39</sup>),

<sup>1</sup>Centre for Polar Observation and Modelling, University of Leeds, Leeds, UK. <sup>2</sup>Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA. <sup>3</sup>Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA. \*e-mail: a.shepherd@leeds.ac.uk



**Fig. 1 | Average annual motion of the Antarctic ice sheet and ice shelves, and of the surrounding sea ice in winter.** The ice sheet is drained by around 30 principal flow units and the sea ice transport is generally northwards, with gyres in the Ross and Weddell seas. Grounded-ice and ice-shelf motion are derived from multiple satellite interferometric synthetic aperture radar data acquired<sup>20</sup> between 2007 and 2009. Ice sheet motion flowlines are superimposed on the MODIS mosaic of Antarctica<sup>142</sup>. Imagery from the NASA MODIS instrument, courtesy NASA NSIDC DAAC. Sea ice motion is the mean of daily gridded Polar Pathfinder radiometry obtained during peak winter (September) of each year in the period<sup>143</sup> 1990 to 2016. Sea ice motion vectors are superimposed on a map of mean sea ice concentration derived from passive microwave brightness temperatures<sup>144</sup> in September 1990 to 2016. Also shown are the average minimum extent of sea ice recorded<sup>144</sup> between 1990 to 2016 (blue dashed boundaries), the grounded ice sheet and the floating ice shelves (black boundaries), and the bathymetry of the surrounding ocean<sup>145</sup>. Active subglacial lakes (light blue) were mapped using satellite radar and laser altimetry<sup>51</sup>.

around much of the continent, with monthly sampling, and over multi-decadal periods (Fig. 2). This allows signals of short-term variability to be separated from longer-term trends. Although most of Antarctica has remained stable over the past 25 years, there are clear patterns of imbalance in many coastal sectors—such as the thickening of the Kamb Ice Stream and the thinning of glaciers flowing into the Amundsen Sea and at the Antarctic Peninsula. These changes reflect imbalance between ice flow and snow accumulation within the surrounding catchments. The pace of ice flow at the Kamb Ice Stream is unusually low<sup>40</sup> and has not altered in recent decades, but analysis of ice-penetrating radar measurements<sup>41</sup> shows that it stagnated over a century ago. Elsewhere, inland glacier thinning is almost exclusively coincident with contemporaneous ice speedup<sup>21,42,43</sup> (indicating that the thinning is dynamic in nature) and with perturbations at the marine termini of the glaciers<sup>44</sup> (indicating that the thinning has resulted from ocean forcing).

### Active subglacial lakes

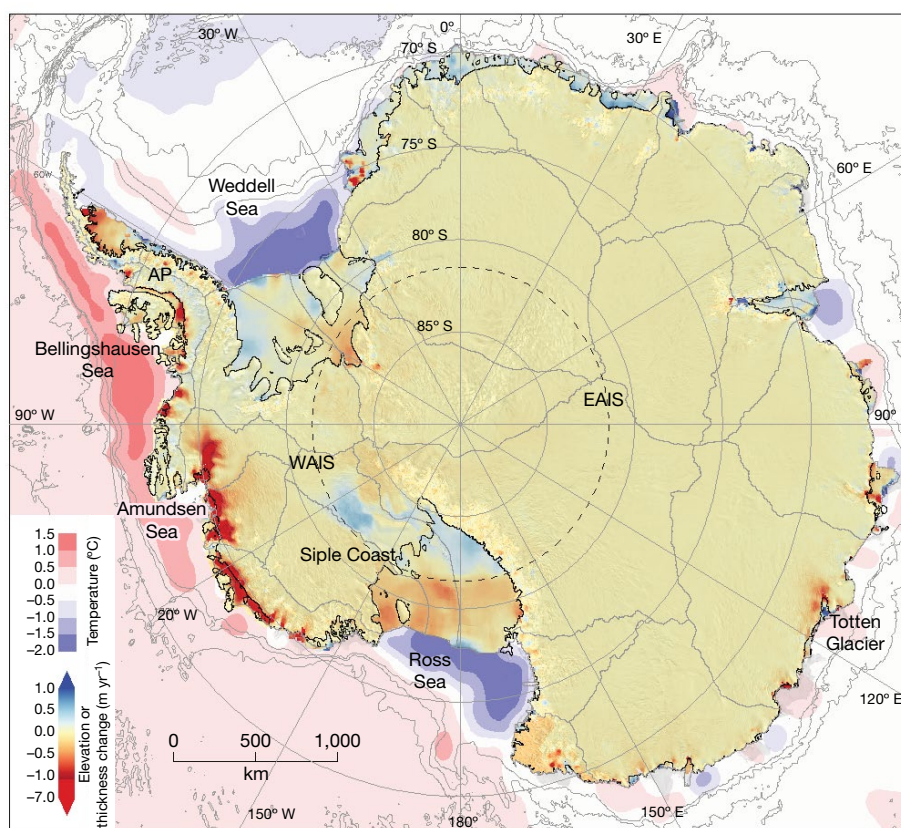
A surprising application of satellite observations has been the monitoring of movement of water beneath the Antarctic Ice Sheet. Over three hundred subglacial lakes—bodies of liquid water at the ice sheet base—have been discovered in Antarctica (Fig. 1) using ice-penetrating radar<sup>45</sup>, and these were at first considered to be isolated and stable reservoirs. However, localized and episodic rises and falls of the ice sheet surface were then spotted in satellite interferometric<sup>46</sup> and altimetric records<sup>47,48</sup>, suggesting otherwise. These fluctuations, amounting to changes in height of 1–10 m over sub-decadal timescales, are interpreted to be the surface expressions of water transferring between active subglacial lake networks. More than a hundred active lakes have now been identified using this approach<sup>49</sup>, and monitoring of their evolution has led to improved understanding of how Antarctic subglacial

water systems evolve, and the consequences of this variability<sup>50</sup>. At the Whillans, Mercer and Recovery ice streams, the Crane and Byrd glaciers, and in eastern Wilkes Land, for example, more than a decade of satellite measurements have been acquired<sup>51</sup>. Thanks to these data, we now know that in addition to periodically flushing subglacial cavities, the presence<sup>52</sup> of and fluctuations<sup>27</sup> in subglacial lake water can lubricate ice flow in parts of the continent.

### Ice shelves

When Antarctic glacier ice reaches the ocean it often remains intact, forming floating ice shelves in sheltered embayments. All together there are more than 300 Antarctic ice shelves, fringing three-quarters of the continent and extending the ice sheet area<sup>4</sup> by some 13%. Their average thicknesses range<sup>53</sup> from 300 m to 2,500 m, and peak at the grounding line, where the shelves are fed by inland glaciers. Ice shelves can provide mechanical support for the grounded ice sheet upstream, through contact with confining side walls or sea mounts<sup>54</sup>. Downstream, they thin as the ice spreads, and they gain and lose additional mass primarily through snow accumulation, iceberg calving and basal ice melting. Basal melting is driven by several processes<sup>5</sup> including the formation of high-salinity water during winter sea ice growth, tidal mixing of seasonally warm water, and the intrusion of warm ocean currents into sub-shelf cavities. Meteorological and oceanographic conditions can also lead to surface melting and basal ice freezing. In some cases, it can take more than a thousand years for ice to travel through Antarctic ice shelves from the grounding line to the calving front<sup>55</sup>, and geological records show that they have been a persistent element of the climate system throughout the Holocene period (for example, see ref. <sup>56</sup>). Their dependence on a wide range of factors makes ice shelves a sensitive indicator of environmental change<sup>57</sup>.





**Fig. 2 | Average trend in the elevation and thickness of Antarctic grounded ice and ice shelves, respectively, determined between 1992 and 2017 north of 81.5° S (dashed grey circle), and between 2010 and 2017 elsewhere.** Also shown is the depth of (from the PANGAEA database in the supplement to ref. <sup>146</sup>) and estimated ocean temperature<sup>147</sup> at the sea floor around the continent. Changes in grounded-ice and ice-shelf thickness were estimated using repeat satellite altimetry following the methods of refs <sup>3,36</sup> and <sup>148</sup>. Thickness trends are superimposed on an optical image mosaic of the floating and grounded ice<sup>142</sup>, and is divided (grey lines) into the principal ice drainage catchments<sup>2</sup>. Since 1992, the grounded

ice sheet and its peripheral ice shelves have thinned in locations adjacent to warm ocean currents. Although the East Antarctic Ice Sheet is mostly stable, there have been marked changes in West Antarctica, including accelerated thinning of glaciers draining the Amundsen Sea sector and constant thickening of glaciers in southerly catchments of the Siple Coast. This thinning is a response to ocean-driven melting of ice shelves at glacier termini<sup>12</sup>, and the thickening is associated with stagnation of ice flow due to a loss of basal lubrication<sup>149</sup>. At the Antarctic Peninsula, ice shelf collapse<sup>14</sup> has triggered inland glacier acceleration<sup>25,68</sup> and thinning<sup>15</sup>. Imagery from the NASA MODIS instrument, courtesy NASA NSIDC DAAC.

### Ice shelf imbalance

Trends in ice shelf area, thickness and flow can be detected using a wide range of satellite sensors, and a host of other properties can be inferred from these measurements. Ice shelf area can be measured using optical and radar satellite imagery, and this has been used, for example, to chart long-term changes in their extent<sup>14</sup>. A series of satellite radar and laser altimeter missions have provided near-continuous observations of ice shelf surface elevation for several decades, and these have formed the basis of ice shelf thickness<sup>58</sup> and thickness change<sup>3,11,59</sup> estimates on the assumption that the ice is buoyant within the surrounding ocean. These estimates require careful treatment of fluctuations in ocean tide<sup>13</sup> and of changes in the firn column thickness<sup>60</sup>. When combined, measurements of ice shelf area and thickness change allow the volume and mass trends of the ice shelves to be derived. Ice shelf flow can be monitored with repeat pass satellite optical<sup>61</sup> and radar<sup>62</sup> imagery and, if contemporaneous changes in both the flow and thickness of ice shelves are available, the rate of steady state<sup>63</sup> and net<sup>12,64,65</sup> basal ice melting can be determined.

Analysis of ice shelf surface elevation measurements derived from multi-mission satellite altimetry (Fig. 2) has allowed the decadal mass change and principal environmental forcing mechanisms of ice shelves to be identified<sup>3,11,59,66</sup>. Although the major Ross, Filchner-Ronne, and Amery ice shelves have remained stable since the 1990s, many ice shelves in West Antarctica have experienced long-term thinning over the same period. In the locations where retreat or thinning

have occurred, the grounded ice inland has also been destabilized. The dominant control on this pattern is believed to be the presence (or absence) of warm ocean currents offshore<sup>59</sup>. Altogether, the volume of Antarctic ice shelves has declined through net overall thinning ( $166 \pm 48 \text{ km}^3 \text{ yr}^{-1}$  between 1994 and 2012; ref. <sup>11</sup>) and through progressive calving-front retreat of those at the Antarctic Peninsula ( $210 \pm 27 \text{ km}^3 \text{ yr}^{-1}$  between 1994 and 2008; ref. <sup>3</sup>). Combined, these losses amount to less than 1% of their volume. However, the highest ice shelf thinning rates have occurred in the Amundsen and Bellingshausen seas<sup>12</sup>, where five have lost between 10% and 18% of their thickness<sup>11</sup> owing to ocean-driven melting at their bases<sup>67</sup>. The effects of the wider El Niño–Southern Oscillation have also been detected alongside these longer-term trends<sup>66</sup>.

### Ice shelf collapse

Ice shelves at the Antarctic Peninsula (Fig. 3) are especially vulnerable, because they are situated at the most northerly latitude on the continent, where temperatures are relatively high and summertime melting is common. In recent decades, a number of these ice shelves have disintegrated in part or entirely<sup>14</sup>. Notable examples include the substantial (>70%) retreat of the Larsen B<sup>15,68</sup> and Wilkins<sup>69</sup> ice shelves, and the effective collapse (>90%) of the Prince Gustav Channel<sup>70</sup>, Larsen A<sup>71</sup>, and Wordie<sup>72</sup> ice shelves. Since the 1950s, the combined area<sup>14,73</sup> of Antarctic ice shelves lost through retreat and collapse has been  $33,917 \text{ km}^2$ , or 22% of their original extent. Analysis of the geological

record<sup>56,74</sup> has confirmed that the collapse events are unique during the Holocene period.

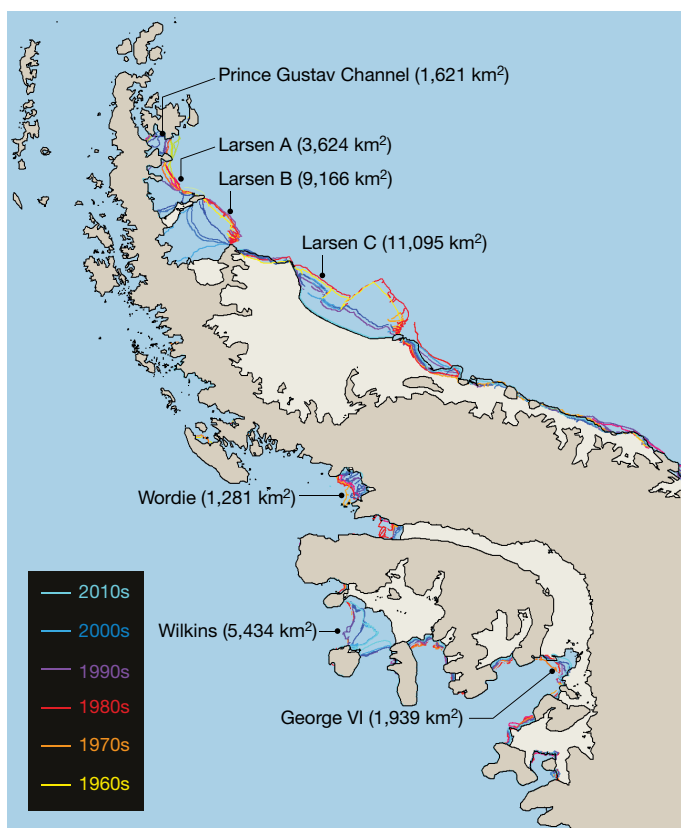
The retreat and collapse of Antarctic Peninsula ice shelves has occurred in tandem with a rapid regional atmospheric warming happening at several times the global trend<sup>57</sup>. These events have been linked; warmer air temperatures lead to intensified surface melting<sup>75</sup>, which is believed to cause hydraulic fracture of surface crevasses followed by ice shelf collapse<sup>76</sup>. Several Antarctic Peninsula ice shelves have also thinned in the decades leading up to their collapse<sup>11,13,59</sup>, primarily through ocean-driven melting at their base. This thinning may contribute to instability by weakening ice shelf lateral margins before fracture<sup>77</sup>, and by enhancing rates of iceberg calving<sup>78</sup>. The relationship is, however, not universal; for example, although the Wilkins ice shelf collapsed in 2009, it did not thin in the preceding five years<sup>79</sup>. Indeed, recent satellite altimetry<sup>80</sup> shows that the surface elevation of the Larsen C ice shelf increased in the preceding decade, in response to cooler (and not warmer) summertime temperatures. And although the observational evidence suggests that stability of Antarctic ice shelves depends also on their thickness, geographical location or setting, the extent to which those shelves that are partially or wholly intact will continue to resist collapse remains uncertain.

The collapse of ice shelves does not contribute directly to sea level rise, because they are afloat. However, there is an indirect effect: observations show that the grounded tributaries to the Larsen A<sup>81</sup> and Larsen B<sup>15,68</sup> ice shelves did speed up in response to the removal of the floating ice, which is presumed to have offered resistance. So far, ice shelf retreat and collapse has been restricted to those shelves situated at the Antarctic Peninsula, in relatively warm climates, and has not threatened those farther south, on the fringes of the East and West Antarctic Ice Sheets. The largest recorded reduction in ice shelf area at the Antarctic Peninsula so far has been the calving of the 11,095 km<sup>2</sup> A-68 tabular iceberg (Fig. 3) from the Larsen C ice shelf<sup>73</sup> in 2016. However, this iceberg represented just 7% of the ice shelf's area, and was similar in size to one (the A-20 iceberg) that broke free<sup>82</sup> in 1986. It is, therefore, not without precedent—even during the relatively short satellite era—and there is as yet no evidence that either breakaway disturbed the remaining ice shelf, or was anything other than routine iceberg production.

### Buttressing of grounded ice

The term 'buttressing' is used to describe the resistive forces imparted to a grounded ice sheet by its peripheral ice shelves. In its absence, rates of ice sheet discharge increase nonlinearly with ice thickness, making the grounding lines of marine-based ice sheets difficult to stabilize because their bedrock tends to deepen inland<sup>83</sup>. Floating ice shelves can, however, exert drag as they flow over and around seamounts (pinning points) or as a result of their lateral confinement, and the extent to which this drag can mitigate unstable retreat has long been the subject of glaciological debate<sup>54</sup>. In recent years, the rapid response of Antarctic glaciers to the collapse and thinning of ice shelves at their termini has led to a reassessment of their resistive properties. At the Antarctic Peninsula, for example, glaciers flowing into the former Larsen A, Larsen B and Wordie ice shelves have surged<sup>24,68,84</sup> after their collapse<sup>14</sup>, as too have ice streams flowing into the Amundsen and Bellingshausen seas<sup>21,43</sup> in the wake of ice shelf thinning<sup>3,12,79</sup> and grounding-line retreat<sup>85,86</sup>. Although the events at the Antarctic Peninsula have been attributed to the destabilizing effect of increased melting at the surface of ice shelves, following regional atmospheric warming<sup>15,68</sup>, the events in the Amundsen and Bellingshausen seas are now firmly linked to enhanced ocean-driven melting at the bases of the ice shelves, owing to the intrusion of warm circumpolar deep water into the cavities beneath them<sup>67,87</sup>.

Although the reservoir of grounded ice at the Antarctic Peninsula is relatively modest<sup>13</sup>, destabilization of the Amundsen Sea sector glaciers is a matter of considerable concern, because the pace of ice drawdown during the satellite era has been swift<sup>88</sup>, and because these glaciers contain enough ice to raise global sea levels by more than a metre<sup>4</sup>. Over the past two decades, for example, surface lowering has spread inland



**Fig. 3 | Temporal changes in the location of ice shelf barriers at the Antarctic Peninsula.** The outlines of the ice shelves (coloured lines) are as determined from satellite imagery since the 1950s, and the net reduction in area over the same period is shown in parentheses<sup>14,73</sup>. Pale (beige) areas are ice shelves, darker (brown) areas are grounded ice. The reduction in area at the Larsen C ice shelf includes the recent calving of the A-68 tabular iceberg. In the 1950s, the total area of Antarctic Peninsula ice shelves was estimated to be 152,246 km<sup>2</sup>. Since then, an area of 33,917 km<sup>2</sup> has been lost during episodic calving events.

across the drainage basins of the Pine Island and Thwaites glaciers at speeds of between 5 and 15 km yr<sup>-1</sup>, and the majority of their catchments are now in a state of dynamical imbalance (they are thinning owing to accelerated flow). This rapid spreading is a consequence of several connected processes<sup>89</sup> (Box 1): ice shelf thinning leads to initial reductions in sidewall and basal traction at glacier termini, which then causes increased strain rates (flow) within the glacier ice upstream, followed by further grounding-line retreat caused by the associated ice thinning—especially in marine-based sectors of the continent. Glacier speedup may also lead to further reductions in sidewall traction through rifting and fracture, and grounding-line retreat can expose more ice to the ocean melting responsible for the initial imbalance.

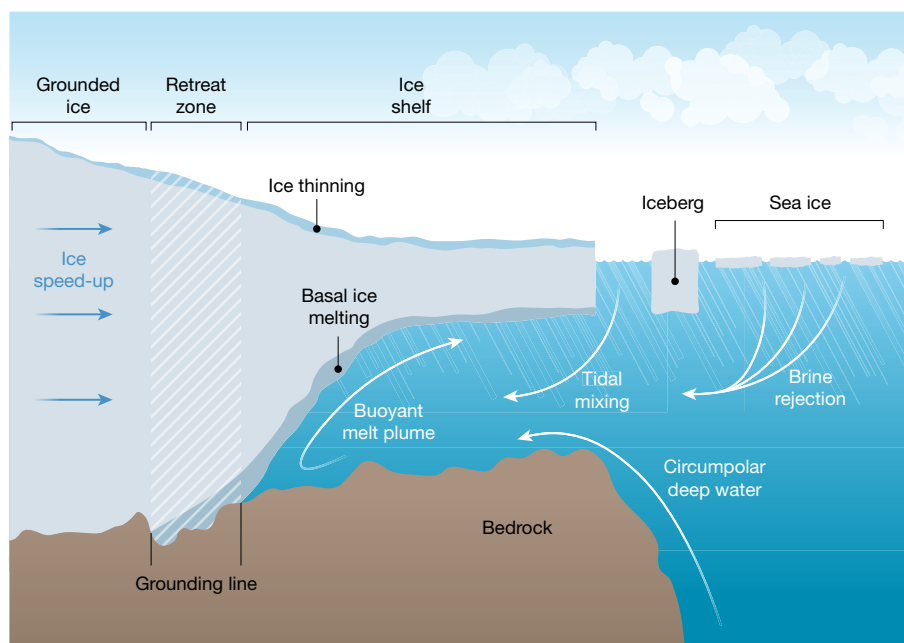
Glaciers flowing into the Amundsen Sea sector of West Antarctica are particularly susceptible to climate forcing, owing to their geometrical configuration and the absence of any substantial ice shelf barrier<sup>90</sup>, and today the pace of ice sheet retreat along parts of this coastline dwarfs that during the Holocene period. The region's ice shelves have thinned<sup>11,12</sup> by 3 to 6 m yr<sup>-1</sup>, and its glacier grounding lines have retreated<sup>85,86</sup> by 10 to 35 km since 1992, which is 20 to 30 times the rate since the Last Glacial Maximum, according to analysis of the marine geological record<sup>91</sup>. In response to these perturbations, the grounded glaciers inland have sped up<sup>21,43</sup> and thinned<sup>32,37</sup> at faster rates. For example, since the early 1990s, rates of ice flow at the Pine Island Glacier terminus have increased<sup>43</sup> by about 1.5 km yr<sup>-1</sup> and rates of ice thinning have risen<sup>88</sup> to over 5 m yr<sup>-1</sup>, and the sector overall contributed 4.5 mm to global sea level rise<sup>38</sup> between 1992 and 2013.

The forcing for these events is now widely regarded to lie in the surrounding ocean, because ice drawdown has originated at and evolved

## Box 1

## Ice shelf buttressing

The figure depicts ice shelf buttressing processes, external forcings and their connectivity. Ice shelves are floating sheets of ice that form as glaciers spread out into the ocean, typically within confined embayments. They are permanently attached to the grounded ice sheet resting on land, and they gain mass through ice flow across the grounding line and local snowfall on the surface, and lose mass by melting at their bases and iceberg calving. If warm water enters the ocean cavity beneath an ice shelf it can drive increased basal ice melting and ice shelf thinning, which in turn leads to retreat of the grounding line—the junction between grounded and floating ice on the sea floor. Ice shelf thinning reduces sidewall (lateral) traction and grounding-line retreat reduces basal traction. Both processes lead to speedup of the grounded ice, which causes grounded ice thinning. Glacier speedup can also lead to weakening of lateral shear margins and increased crevassing. Iceberg calving can also lead to reduced sidewall traction. Sea ice (frozen sea water) can play a part through brine rejection, which drives the production of warm high-salinity shelf water, and by providing an additional buffer that effectively increases buttressing.



from the terminus of neighbouring but distinct ice flow units<sup>42</sup>, and because warm<sup>67</sup> and warming<sup>44</sup> water is present within the cavities beneath their peripheral ice shelves. According to numerical simulations, the Pine Island and Thwaites glaciers<sup>92,93</sup> may contribute a further 4 mm or so to global sea levels over the twenty-first century in response to continued forcing, and it has been concluded<sup>94</sup> that the region is now undergoing marine ice sheet instability, with no geometrical obstacles to prevent irreversible decline. However, satellite observations have revealed that retreat of the Pine Island Glacier halted around 2011<sup>95,96</sup>, and that ice thinning inland abated in the following years<sup>88</sup>. This suggests that the situation is more complicated than a consideration of the glacier geometry alone, and may involve changes in the degree of ocean forcing, as has occurred in the recent past<sup>87,97</sup>.

## Sea ice

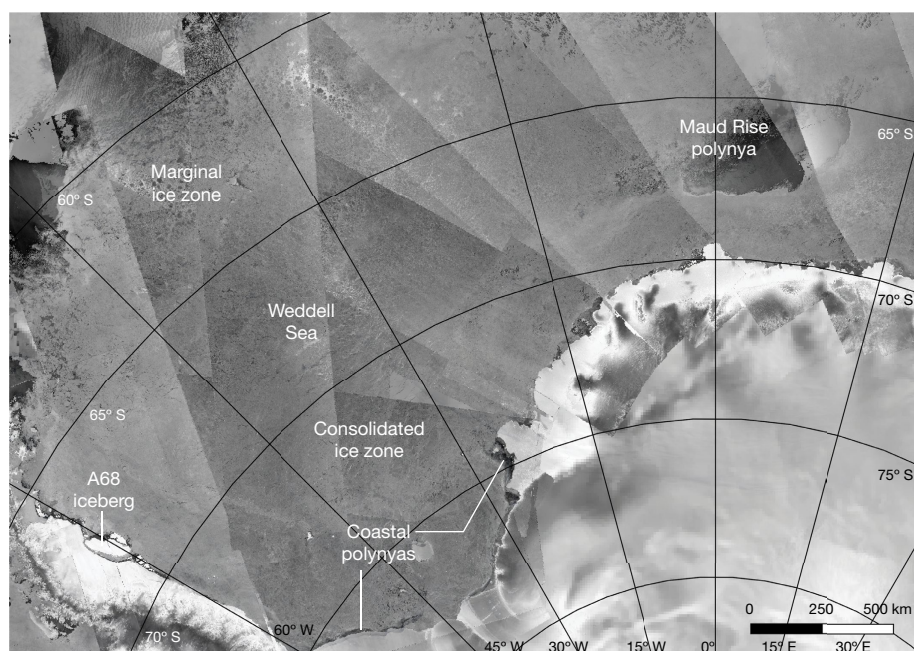
Antarctic sea ice forms as the Southern Ocean surface freezes, and it interacts with the neighbouring ice shelves and grounded ice in many ways (for example, Box 1). Satellite observations (for example, Fig. 1) have allowed us to map the extent<sup>8,98</sup>, thickness<sup>99,100</sup> and drift<sup>101,102</sup> of the sea ice, providing insight into the effects it has on climate<sup>6</sup> and ecosystems<sup>103</sup>. In winter, Antarctic sea ice extends from an inner zone of consolidated pack ice surrounding the continent, to the marginal ice zone near the powerful Antarctic Circumpolar Current, where floes are less concentrated (Figs. 1 and 4). In summer, the sea ice pack retreats to isolated pockets fringing the continent. As it forms, Antarctic sea ice produces high-salinity shelf water when brine is rejected, which then sinks to the seabed. This water drives

buoyant plumes within the cavities beneath floating ice shelves, which melt glacier ice at the grounding line, before returning to the open ocean along their base. Antarctic sea ice is also characterized by local polynyas—persistent gaps in the ice cover (for example, Fig. 4) that are sustained by upwelling warm water, winds, tides and ocean currents. These polynyas are a source of bottom water (dense water occupying depths typically below 4,000 m), and provide a link between the ocean and atmosphere that affects weather and wildlife<sup>103</sup>. Landfast sea ice can also act to stabilize ice shelves and glacier tongues<sup>104</sup>, and to suppress<sup>105</sup> or—upon its breakup—enable<sup>106</sup> iceberg calving.

## Sea ice extent and drift

Fluctuations in the area of Antarctic sea ice have been routinely charted since the late 1970s using passive microwave satellite imagery<sup>8</sup>. Annually, its average extent ranges from  $3.1 \times 10^6$  km<sup>2</sup> in February to  $18.5 \times 10^6$  km<sup>2</sup> in September<sup>1</sup> (for example, Fig. 1). In contrast to the Arctic, where the area of sea ice has declined progressively<sup>1</sup>, there has been a small, positive increase ( $1.6 \pm 0.4\%$  per decade between 1979–2016) in the hemispheric sea ice extent of the Southern Ocean<sup>9</sup>. This trend runs counter to the projections of most climate models<sup>107</sup>, and has occurred alongside a slow warming ( $0.02^\circ\text{C}$  per decade since the 1950s) of the Southern Ocean<sup>108</sup>. Despite the trend, there is some evidence of longer-term decline: reanalysis of early satellite records<sup>109,110</sup> and historical whale catch positions<sup>111</sup> suggest there may have been more ice cover in the 1960s and early 1970s than there is today. In recent years, however, extreme changes have occurred—the extent of Antarctic sea ice reached record maxima in three successive winters





**Fig. 4 | Sea ice in the Weddell Sea in early November 2017, based on a composite of Sentinel-1 synthetic aperture radar imagery, MODIS optical satellite imagery and ASCAT scatterometer data.** The satellite data were acquired during the austral winter when the ice cover is close to its maximum extent, stretching beyond the tip of the Antarctic Peninsula into Drake Passage. The composite reveals details of the diffuse ice cover in the marginal ice zone along the boundary with the open ocean, and the more compact, consolidated ice cover farther south. A large (approximately 35,000 km<sup>2</sup>) ice-free area is visible near Maud Rise (66° S, 5° E); this polynya is formed by thermally driven convection in the

water column, owing to the interaction of ocean currents and seafloor topography. Coastal polynyas are visible along the edges of the Filchner–Ronne, Brunt and Stancombe–Willis ice shelves, where openings in the ice cover are driven by katabatic wind, tides and ocean currents. New sea ice, which rejects brine during formation, is continually produced in these polynyas, making polynyas a source of dense, high-salinity deep water, which plays an important part in the global thermohaline circulation. Also visible is the large, tabular A68 iceberg (located at about 68° S, 60.5° W), which calved from the Larsen C Ice Shelf on 12 July 2017, and is now adrift in the western Weddell Sea.

(2012, 2013 and 2014; ref. <sup>112</sup>), followed by a record summertime minimum<sup>113,114</sup> in March 2017.

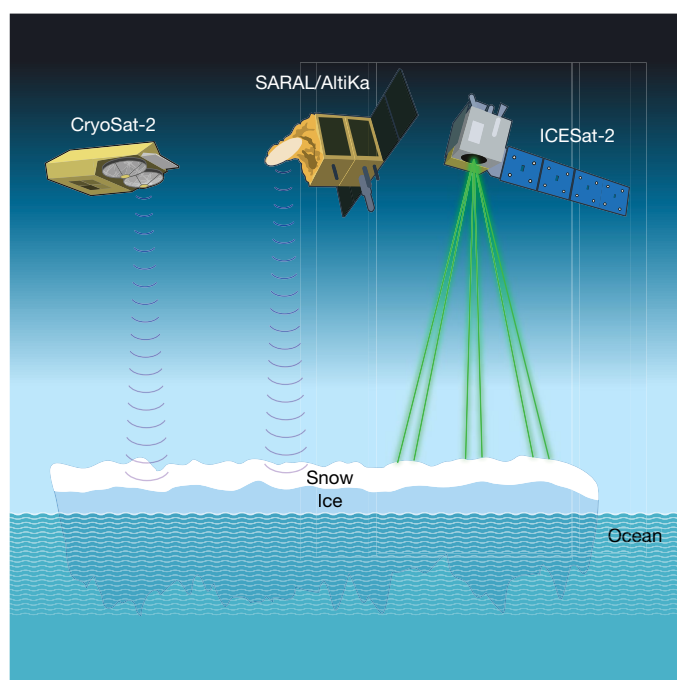
Changes in both atmospheric and oceanic forcing affect the extent of Antarctic sea ice<sup>115,116</sup>, and although total hemispheric extent has shown little overall change, there have been considerable regional variations<sup>117,118</sup>. Although the Weddell Sea, Indian Ocean and the western Pacific Ocean have all seen modest trends ( $1.7\% \pm 0.8\%$ ,  $1.7\% \pm 0.99\%$  and  $1.8\% \pm 1.2\%$  per decade, respectively) in sea ice extent during the satellite era (1979–2016), there have been more substantial trends ( $3.3\% \pm 0.9\%$  and  $-2.9\% \pm 1.4\%$  per decade, respectively) in the Ross Sea and the Amundsen and Bellingshausen seas<sup>9</sup>. The periods during which the western Ross and Bellingshausen seas are ice-free in summer have also changed, decreasing and increasing by two and three months, respectively<sup>118</sup>, between 1979 and 2011. Seasonal and decadal trends in the Weddell and Ross seas (positive) and in the Amundsen and Bellingshausen seas and the western Pacific Ocean (negative) reflect the influence of atmospheric forcing<sup>119</sup>. Although these fluctuations in sea ice extent are strongly correlated with the dominant modes of Southern Hemisphere climate variability<sup>117,119</sup>, other factors are involved<sup>120</sup>. A range of mechanisms have been explored, including changes in oceanic variability<sup>121</sup>, atmospheric circulation<sup>122,123</sup>, stratospheric ozone depletion<sup>124</sup>, meridional wind forcing<sup>102</sup> and freshwater input from ice shelf melt<sup>125,126</sup>.

Understanding the role sea ice plays in the Antarctic climate system also requires a consideration of its dynamics<sup>127</sup>. In the Southern Ocean, sea ice drifts northwards and diverges under the influence of winds and ocean currents (Fig. 1), and the fraction of open water is higher than in the Arctic<sup>128</sup>. Satellite observations have illuminated both local-scale<sup>129,130</sup> and hemisphere-wide<sup>101,102,131</sup> Antarctic sea ice dynamics. Strong, circumpolar, westerly winds drive sea ice eastwards in the outer zonal band, while a nearly continuous westward circumpolar flow exists along the coastal boundary<sup>101</sup>. Persistent atmospheric lows centred at the boundaries of major ocean basins

are the dominant drivers of sea ice motion, and these sustain large-scale gyres in the Weddell and Ross seas<sup>132</sup>. The speed of sea ice drift in the eastern Weddell and Ross seas has increased, in contrast to the western Weddell Sea, where it has decreased<sup>102,132</sup>, though these signals are still small compared to the interannual variability<sup>101</sup>. The general northward trajectory of the Antarctic sea ice pack also affects its age and thickness; rarely does it survive for more than two years, and the average thickness of floes (typically in the range 0.6–1.2 m) and pressure ridges are smaller than in the Arctic Ocean<sup>7</sup>. Locally, katabatic winds, tides and ocean currents sustain coastal polynyas through sea ice drift around the continent (for example, Fig. 4), providing a link between the sea ice pack and the ice sheet through their initiation of plumes beneath floating ice shelves<sup>133</sup>.

## Summary and outlook

In just three decades, satellites have transformed our appreciation of the extent and pace of change in the Antarctic cryosphere. Despite being remote, fluctuations in its ice cover have a global impact. The continent holds Earth's primary freshwater reservoir<sup>4</sup> and, together with its surrounding ice shelves<sup>3</sup> and sea ice<sup>1</sup>, blankets 6% of the planet in ice during the austral winter. Although persistent ice shelves have fringed Antarctica for thousands of years<sup>36</sup>, there is now widespread evidence of changes in their extent<sup>14</sup> and thickness<sup>3,11</sup>. Altogether, their volume has decreased by more than 300 km<sup>3</sup> yr<sup>-1</sup> since 1994<sup>3,11</sup>, notably due to collapse and calving at the Antarctic Peninsula and rapid thinning of those in the Amundsen and Bellingshausen seas. These events have triggered retreat<sup>85,86</sup> and acceleration<sup>21,43</sup> of marine-terminating glaciers and ice streams around the continent, leading to the drawdown of ice from their inland catchments<sup>39,42</sup>. Since 1992, the grounded ice sheet has lost  $1,350 \pm 1,010$  Gt of ice, causing a net  $3.8 \pm 2.8$  mm contribution to global sea level rise<sup>36</sup>. The waxing and waning of Antarctic sea ice influences the planetary albedo, oceanic circulation, marine productivity and ecosystems<sup>6,103</sup>. Although its extent has increased by  $1.6\% \pm 0.4\%$  per



**Fig. 5 | Schematic of a sea ice floe as observed by CryoSat-2 and AltiKa.** Floe thicknesses are typically derived from measurements of their freeboard (the portion protruding above the ocean surface), an estimate of the snow loading and the principle of buoyancy<sup>99</sup>. CryoSat-2 and AltiKa operate at different radar frequencies, and their echoes scatter from locations near to the lower and upper boundaries of the snow layer, respectively<sup>139,140</sup>. Measurements acquired at both frequencies could provide a direct measurement of the snow loading, improving the certainty of sea ice floe thickness estimates.

decade since 1979<sup>9</sup>, there are large regional variations<sup>9,117,118</sup>, and there is evidence from historical records<sup>109–111</sup> of a longer-term decline. These discoveries, and many more, have transformed our understanding of the state of Antarctic ice.

Even though considerable progress has been made during the satellite era, particularly on understanding the ice sheet, key questions remain unanswered. For example, the detailed pattern of glacier change at the Antarctic Peninsula is not well known, because the rugged terrain poses a challenge for traditional remote sensing methods. Though modest, the mass balance of the East Antarctic Ice Sheet nevertheless remains uncertain, because its detection is complicated by uncertainties in rates of snowfall and glacial isostatic adjustment. And the evolution and impacts of abrupt subglacial lake drainage events is poorly defined, because frequent measurements of ice elevation and flow changes are often lacking at the local scale. But understanding the thickness of sea ice across the Southern Hemisphere and the nature of ice shelf collapse and retreat are pressing concerns. Even though the range of parameters that can now be measured on grounded ice and on ice shelves may be considered comprehensive, available satellite observations are of insufficient spatial and temporal sampling to fully understand the nature and evolution of the processes that are driving contemporary imbalance. Although the satellite altimeter record has been used to resolve Southern Ocean dynamics, determining sea ice thickness—a key measure of its volume and longevity—from measurements of its freeboard (the portion protruding above the ocean surface) are hampered by poor knowledge of snow loading and its impact on the satellite retrieval.

In the case of Antarctic sea ice, uncertainties in the degree of radar penetration into the snowpack<sup>133,134</sup> has so far limited the use of the 25-year radar altimeter record for measuring its thickness. Some advances have been made using laser altimetry<sup>99,100</sup>, which scatters from the surface of the overlying snow, but continental-scale trends in Antarctic sea ice thickness and volume nevertheless remain elusive

owing to the paucity of in situ measurements. One way to tackle this problem is to exploit the relationship between the amount and roughness of snow on sea ice and its total thickness<sup>136</sup>, an approach that may be realized with the launch of ICESat-2, which has a laser capable of detecting surface roughness and thickness<sup>137,138</sup>. Another possibility is to combine freeboard measurements retrieved from different scattering horizons to estimate the snow load directly, for example using observations acquired by the CryoSat-2 Ku-band and AltiKa Ka-band radar altimeters<sup>139,140</sup> (Fig. 5) and, in the future, ICESat-2. New techniques are also emerging that will enable us to map the extent, type, age, drift and roughness of sea ice with fine resolution using synthetic aperture radar imagery.

Over land ice, the record of ice sheet motion data are too sparse to determine whether changes in flow have occurred on sub-annual timescales across much of the continent. On this point, the outlook is promising thanks to the systematic acquisition plans of the Sentinel-1 synthetic aperture radar and Landsat-8 optical imager missions. A key unanswered science question is how long it will take for the ice shelves that are currently thinning to reach a point whereby they are no longer providing effective buttressing for the grounded ice inland. To address this, observations are required with sufficient frequency to track the events themselves, which, in the case of ice shelf collapses, have taken place over months or even days<sup>14</sup>. Although it is possible to monitor grounding-line migration with high precision using synthetic aperture radar interferometry<sup>85,86</sup>, the revisit period of satellite missions is currently too long for the technique to be effective over rapidly deforming ice, and so other methods—such as repeat satellite altimetry<sup>96,141</sup>—will need to be exploited to track this precursor to ice sheet dynamical imbalance.

The past decade has been a golden era for satellite glaciology, with a host of different sensors in orbit simultaneously. However, measuring ice loss from Antarctica at the continental scale is today heavily reliant upon a single ageing mission—CryoSat-2—which, at 8 years old, has now more than doubled its planned lifetime, and the continuity of passive microwave observations of sea ice concentration and extent remains uncertain. Given the societal importance of changes in ice cover and global sea level, support for current and planned observational platforms should remain a central goal of geoscience research—a goal that can only be achieved by concerted, long-term international collaboration.

Received: 8 December 2017; Accepted: 21 March 2018;

Published online 13 June 2018.

1. Parkinson, C. L. Global sea ice coverage from satellite data: annual cycle and 35-yr trends. *J. Clim.* **27**, 9377–9382 (2014).  
**This paper is a recent assessment of multi-decadal trends in global sea ice extent as derived from satellite passive microwave radiometer data, confirming that the losses in ice extent in the Arctic Ocean far exceed gains in the Southern Ocean.**
2. Zwally, H. J., Giovinetto, M. B., Beckley, M. A. & Saba, J. L. *Antarctic and Greenland Drainage Systems* (GSFC Cryospheric Sciences Laboratory, 2012).
3. Shepherd, A. et al. Recent loss of floating ice and the consequent sea level contribution. *Geophys. Res. Lett.* **37**, <https://doi.org/10.1029/2010GL042496> (2010).
4. Fretwell, P. et al. Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *Cryosphere* **7**, 375–393 (2013).  
**This paper presents models of the Antarctic ice sheet and ice shelf thickness, determined from a compilation of airborne and satellite remote sensing, that are widely used across the glaciological community and beyond.**
5. Jacobs, S. S., Helmer, H. H., Doake, C. S. M., Jenkins, A. & Frolich, R. M. Melting of ice shelves and the mass balance of Antarctica. *J. Glaciol.* **38**, 375–387 (1992).
6. Massom, R. A. & Stammerjohn, S. E. Antarctic sea ice change and variability—physical and ecological implications. *Polar Sci.* **4**, 149–186 (2010).
7. Worby, A. P. et al. Thickness distribution of Antarctic sea ice. *J. Geophys. Res. Oceans* **113**, <https://doi.org/10.1029/2007JC004254> (2008).
8. Zwally, H. J., Parkinson, C. L. & Comiso, J. C. Variability of Antarctic sea ice and changes in carbon dioxide. *Science* **220**, 1005–1012 (1983).  
**As an early application of satellite radar imagery for tracking trends in the extent of sea ice in the Southern Hemisphere, this paper is a seminal study.**
9. De Santis, A., Maier, E., Gomez, R. & Gonzalez, I. Antarctica, 1979–2016 sea ice extent: total versus regional trends, anomalies, and correlation with climatological variables. *Int. J. Remote Sens.* **38**, 7566–7584 (2017).



10. The IMBIE Team. Mass balance of the Antarctic ice sheet from 1992 to 2017. *Nature* **558**, <https://doi.org/10.1038/s41586-018-0179-y> (2018).  
**This large collaborative work presents an updated comparison and synthesis of many individual estimates of Antarctic ice sheet mass balance derived from satellite observations to deliver a single result for use by the wider scientific community.**
11. Paolo, F. S., Fricker, H. A. & Padman, L. Volume loss from Antarctic ice shelves is accelerating. *Science* **348**, 327–331 (2015).  
**A multi-mission record (1994 to 2012) of ice-shelf surface height from satellite radar altimetry showed accelerated loss of volume of Antarctica's ice shelves, with early increases in East Antarctica, probably due to accumulation, and substantial losses in West Antarctica, where some ice shelves thinned by up to 18% over the last 18 years.**
12. Shepherd, A., Wingham, D. & Rignot, E. Warm ocean is eroding West Antarctic Ice Sheet. *Geophys. Res. Lett.* **31**, 1–4 (2004).
13. Shepherd, A., Wingham, D., Payne, T. & Skvarca, P. Larsen Ice Shelf has progressively thinned. *Science* **302**, 856–859 (2003).  
**This paper describes the first application of satellite measurements for detecting trends in the thickness of Antarctic ice shelves, providing direct observations of contemporary imbalance and evidence that ocean-driven melting is a destabilizing force.**
14. Cook, A. J. & Vaughan, D. G. Overview of areal changes of the ice shelves on the Antarctic Peninsula over the past 50 years. *Cryosphere* **4**, 77–98 (2010).
15. Scambos, T. A., Bohlander, J. A., Shuman, C. A. & Skvarca, P. Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. *Geophys. Res. Lett.* **31**, L18402 (2004).
16. Drewry, D. J. *Antarctica: Glaciological and Geophysical Folio* (ed. Drewry, D. J.) (Scott Polar Research Institute, University of Cambridge, Cambridge, 1983).
17. Bamber, J. L., Vaughan, D. G. & Joughin, I. Widespread complex flow in the interior of the Antarctic Ice Sheet. *Science* **287**, 1248–1250 (2000).  
**This study was the first to apply the balance-velocity technique to map the continental pattern of ice flow, revealing the intricate nature of the ice sheet glaciers.**
18. Scambos, T. A., Dutkiewicz, M. J., Wilson, J. C. & Bindshadler, R. A. Application of image cross-correlation to the measurement of glacier velocity using satellite image data. *Remote Sens. Environ.* **42**, 177–186 (1992).
19. Goldstein, R. M., Engelhardt, H., Kamb, B. & Frolich, R. M. Satellite radar interferometry for monitoring ice sheet motion: application to an Antarctic ice stream. *Science* **262**, 1525–1530 (1993).  
**This ground-breaking study was the first to explain how the innovative technique of satellite radar interferometry could be applied to glaciology, introducing methods for tracking glacier topography and motion, and the location of ice stream grounding lines.**
20. Rignot, E., Mouginot, J. & Scheuchl, B. Ice flow of the Antarctic Ice Sheet. *Science* **333**, 1427–1430 (2011).
21. Joughin, I., Rignot, E., Rosanova, C. E., Lucchitta, B. K. & Bohlander, J. Timing of recent accelerations of Pine Island Glacier, Antarctica. *Geophys. Res. Lett.* **30**, <https://doi.org/10.1029/2003GL017609> (2003).
22. Rignot, E. Evidence for rapid retreat and mass loss of Thwaites Glacier, West Antarctica. *J. Glaciol.* **47**, 213–222 (2001).
23. Joughin, I., Tulaczyk, S., Bindshadler, R. & Price, S. F. Changes in west Antarctic ice stream velocities: observation and analysis. *J. Geophys. Res. Solid Earth* **107**, <https://doi.org/10.1029/2001JB001029> (2002).
24. Rignot, E. et al. Recent ice loss from the Fleming and other glaciers, Wordie Bay, West Antarctic Peninsula. *Geophys. Res. Lett.* **32**, <https://doi.org/10.1029/2004GL021947> (2005).
25. Rott, H., Müller, F., Nagler, T. & Floricioiu, D. The imbalance of glaciers after disintegration of Larsen-B ice shelf, Antarctic Peninsula. *Cryosphere* **5**, 125–134 (2011).
26. Hogg, A. E. et al. Increased ice flow in Western Palmer Land linked to ocean melting. *Geophys. Res. Lett.* **44**, 4159–4167 (2017).
27. Stearns, L. A., Smith, B. E. & Hamilton, G. S. Increased flow speed on a large east Antarctic outlet glacier caused by subglacial floods. *Nat. Geosci.* **1**, 827–831 (2008).
28. Li, X., Rignot, E., Morlighem, M., Mouginot, J. & Scheuchl, B. Grounding line retreat of Totten Glacier, East Antarctica, 1996 to 2013. *Geophys. Res. Lett.* **42**, 8049–8056 (2015).
29. Rignot, E. et al. Recent Antarctic ice mass loss from radar interferometry and regional climate modelling. *Nat. Geosci.* **1**, 106–110 (2008).
30. Zwally, H. J., Brenner, A. C., Major, J. A., Bindshadler, R. A. & Marsh, J. G. Growth of Greenland ice sheet: measurement. *Science* **246**, 1587–1589 (1989).
31. Wingham, D. J., Ridout, A. J., Scharroo, R., Arthern, R. J. & Shum, C. K. Antarctic elevation change from 1992 to 1996. *Science* **282**, 456–458 (1998).
32. Pritchard, H. D., Arthern, R. J., Vaughan, D. G. & Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461**, 971–975 (2009).
33. Velicogna, I. & Wahr, J. Measurements of time-variable gravity show mass loss in Antarctica. *Science* **311**, 1754–1756 (2006).
34. Luthcke, S. B. et al. Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *J. Glaciol.* **59**, 613–631 (2013).
35. Briggs, K. et al. Charting ice-sheet contributions to global sea-level rise. *Eos* **97**, <https://doi.org/10.1029/2016EO055719> (2016).
36. Shepherd, A. et al. A reconciled estimate of ice-sheet mass balance. *Science* **338**, 1183–1189 (2012).
37. Wingham, D. J., Wallis, D. W. & Shepherd, A. Spatial and temporal evolution of Pine Island Glacier thinning, 1995–2006. *Geophys. Res. Lett.* **36**, <https://doi.org/10.1029/2009GL039126> (2009).
38. Sutterley, T. C. et al. Mass loss of the Amundsen Sea Embayment of West Antarctica from four independent techniques. *Geophys. Res. Lett.* **41**, 8421–8428 (2014).
39. Shepherd, A., Wingham, D. J., Mansley, J. A. D. & Corr, H. F. J. Inland thinning of Pine Island Glacier, West Antarctica. *Science* **291**, 862–864 (2001).
40. Whillans, I. M., Bolzan, J. & Shabtaie, S. Velocity of ice streams B and C, Antarctica. *J. Geophys. Res.* **92**, 8895–8902 (1987).
41. Retzlaff, R. & Bentley, C. R. Timing of stagnation of ice stream C, West Antarctica, from short-pulse radar studies of buried surface crevasses. *J. Glaciol.* **39**, 553–561 (1993).
42. Shepherd, A., Wingham, D. J. & Mansley, J. A. D. Inland thinning of the Amundsen Sea sector, West Antarctica. *Geophys. Res. Lett.* **29**, <https://doi.org/10.1029/2001GL014183> (2002).
43. Mouginot, J., Rignot, E. & Scheuchl, B. Sustained increase in ice discharge from the Amundsen Sea Embayment, West Antarctica, from 1973 to 2013. *Geophys. Res. Lett.* **41**, 1576–1584 (2014).
44. Jacobs, S. S., Jenkins, A., Giulivi, C. F. & Dutrieux, P. Stronger ocean circulation and increased melting under Pine Island Glacier ice shelf. *Nat. Geosci.* **4**, 519–523 (2011).
45. Siegert, M. J., Carter, S., Tabacco, I., Popov, S. & Blankenship, D. D. A revised inventory of Antarctic subglacial lakes. *Antarct. Sci.* **17**, 453–460 (2005).
46. Gray, L. et al. Evidence for subglacial water transport in the West Antarctic Ice Sheet through three-dimensional satellite radar interferometry. *Geophys. Res. Lett.* **32**, <https://doi.org/10.1029/2004GL021387> (2005).  
**This study was the first to detect the surface expression of water transport beneath the Antarctic ice sheet, a new approach for studying the hydrology of the continent's subglacial lakes.**
47. Wingham, D. J., Siegert, M. J., Shepherd, A. & Muir, A. S. Rapid discharge connects Antarctic subglacial lakes. *Nature* **440**, 1033–1036 (2006).
48. Fricker, H. A., Scambos, T., Bindshadler, R. & Padman, L. An active subglacial water system in West Antarctica mapped from space. *Science* **315**, 1544–1548 (2007).
49. Smith, B. E., Helen, A. F., Ian, R. J. & Tulaczyk, S. An inventory of active subglacial lakes in Antarctica detected by ICESat (2003–2008). *J. Glaciol.* **55**, 573–595 (2009).
50. Bell, R. E. The role of subglacial water in ice-sheet mass balance. *Nat. Geosci.* **1**, 297–304 (2008).
51. Siegfried, M. R. & Fricker, H. Thirteen years of subglacial lake activity in Antarctica from multi-mission satellite altimetry. *Ann. Glaciol.* <https://doi.org/10.1017/aog.2017.36> (2018).
52. Bell, R. E., Studinger, M., Shuman, C. A., Fahnestock, M. A. & Joughin, I. Large subglacial lakes in East Antarctica at the onset of fast-flowing ice streams. *Nature* **445**, 904–907 (2007).
53. Schaffer, J. et al. A global, high-resolution data set of ice sheet topography, cavity geometry, and ocean bathymetry. *Earth Syst. Sci. Data* **8**, 543–557 (2016).
54. Weertman, J. Stability of the junction of an ice sheet and an ice shelf. *J. Glaciol.* **13**, 3–11 (1974).
55. Fahnestock, M. A., Scambos, T. A., Bindshadler, R. A. & Kvaran, G. A millennium of variable ice flow recorded by the Ross ice shelf, Antarctica. *J. Glaciol.* **46**, 652–664 (2000).
56. Domack, E. et al. Stability of the Larsen B ice shelf on the Antarctic Peninsula during the Holocene epoch. *Nature* **436**, 681–685 (2005).
57. Vaughan, D. G. et al. Recent rapid regional climate warming on the Antarctic Peninsula. *Clim. Change* **60**, 243–274 (2003).
58. Griggs, J. A. & Bamber, J. L. Antarctic ice-shelf thickness from satellite radar altimetry. *J. Glaciol.* **57**, 485–498 (2011).
59. Pritchard, H. D. et al. Antarctic ice-sheet loss driven by basal melting of ice shelves. *Nature* **484**, 502–505 (2012).
60. Helsen, M. M. et al. Elevation changes in Antarctica mainly determined by accumulation variability. *Science* **320**, 1626–1629 (2008).
61. Rosanova, C. E., Lucchitta, B. K. & Ferrigno, J. G. Velocities of Thwaites Glacier and smaller glaciers along the Marie Byrd Land coast, West Antarctica. *Ann. Glaciol.* **27**, 47–53 (1998).
62. Rack, W., Doake, C. S. M., Rott, H., Siegel, A. & Skvarca, P. Interferometric analysis of the deformation pattern of the northern Larsen Ice Shelf, Antarctic Peninsula, compared to field measurement and numerical modeling. *Ann. Glaciol.* **31**, 205–210 (2000).
63. Rignot, E. & Jacobs, S. S. Rapid bottom melting widespread near Antarctic Ice Sheet grounding lines. *Science* **296**, 2020–2023 (2002).  
**In reporting satellite-derived estimates of ice shelf basal melting, this study was among the first to assess ice-ocean interactions and to highlight regional variations in ocean forcing.**
64. Rignot, E., Jacobs, S., Mouginot, J. & Scheuchl, B. Ice-shelf melting around Antarctica. *Science* **341**, 266–270 (2013).
65. Depoorter, M. A. et al. Calving fluxes and basal melt rates of Antarctic ice shelves. *Nature* **502**, 89–92 (2013).
66. Paolo, F. S. et al. Response of Pacific-sector Antarctic ice shelves to the El Niño/Southern Oscillation. *Nat. Geosci.* **11**, 121–126 (2018).
67. Jacobs, S. S., Hellmer, H. H. & Jenkins, A. Antarctic Ice Sheet melting in the southeast Pacific. *Geophys. Res. Lett.* **23**, 957–960 (1996).
68. Rignot, E. et al. Accelerated ice discharge from the Antarctic Peninsula following the collapse of Larsen B ice shelf. *Geophys. Res. Lett.* **31**, L18401 (2004).
69. Humbert, A. & Braun, M. The Wilkins Ice Shelf, Antarctica: break-up along failure zones. *J. Glaciol.* **54**, 943–944 (2008).
70. Cooper, A. P. R. Historical observations of Prince Gustav ice shelf. *Polar Rec.* **33**, 285–294 (1997).



71. Skvarca, P. Fast recession of the northern Larsen Ice Shelf monitored by space images. *Ann. Glaciol.* **17**, 317–321 (1993).
72. Doake, C. S. M. & Vaughan, D. G. Rapid disintegration of the Wordie Ice Shelf in response to atmospheric warming. *Nature* **350**, 328–330 (1991).
73. Hogg, A. E. & Gudmundsson, G. H. Impacts of the Larsen-C Ice Shelf calving event. *Nat. Clim. Change* **7**, 540–542 (2017).
74. Pudsey, C. J. & Evans, J. First survey of Antarctic sub-ice shelf sediments reveals mid-Holocene ice shelf retreat. *Geology* **29**, 787–790 (2001).
75. van den Broeke, M. Strong surface melting preceded collapse of Antarctic Peninsula ice shelf. *Geophys. Res. Lett.* **32**, https://doi.org/10.1029/2005GL023247 (2005).
76. Scambos, T., Hulbe, C. & Fahnestock, M. A. Climate-induced ice shelf disintegration in the Antarctic Peninsula. *Antarct. Res. Ser.* **76**, 335–347 (2003).
77. Vieli, A., Payne, A. J., Shepherd, A. & Du, Z. Causes of pre-collapse changes of the Larsen B ice shelf: numerical modelling and assimilation of satellite observations. *Earth Planet. Sci. Lett.* **259**, 297–306 (2007).
78. Liu, Y. et al. Ocean-driven thinning enhances iceberg calving and retreat of Antarctic ice shelves. *Proc. Natl Acad. Sci. USA* **112**, 3263–3268 (2015).
79. Fricker, H. A. & Padman, L. Thirty years of elevation change on Antarctic Peninsula ice shelves from multitemporal satellite radar altimetry. *J. Geophys. Res. Oceans* **117**, https://doi.org/10.1029/2011JC007126 (2012).
80. Adusumilli, S. et al. Variable basal melt rates of Antarctic Peninsula ice shelves, 1994–2016. *Geophys. Res. Lett.* https://doi.org/10.1002/2017GL076652 (in the press).
81. Royston, S. & Gudmundsson, G. H. Changes in ice-shelf buttressing following the collapse of Larsen A Ice Shelf, Antarctica, and the resulting impact on tributaries. *J. Glaciol.* **62**, 905–911 (2016).
82. Phillips, H. A. & Laxon, S. W. Tracking of Antarctic tabular icebergs using passive microwave radiometry. *Int. J. Remote Sens.* **16**, 399–405 (1995).  
**By tracking a large tabular iceberg that calved from Larsen C Ice Shelf with passive microwave imagery, this paper demonstrated how satellite imagery can be used to detect the calving of large, tabular icebergs from Antarctica, and to chart their motion as they drift around the continent.**
83. Schoof, C. Ice sheet grounding line dynamics: steady states, stability, and hysteresis. *J. Geophys. Res. Earth Surf.* **112**, https://doi.org/10.1029/2006JF000664 (2007).
84. De Angelis, H. & Skvarca, P. Glacier surge after ice shelf collapse. *Science* **299**, 1560–1562 (2003).  
**Although qualitative in nature, this paper was the first to confirm that the disintegration of the Larsen ice shelf triggered increase flow of the grounded ice upstream, by tracking glacial geomorphological features in airborne and satellite imagery.**
85. Rignot, E. J. Fast recession of a West Antarctic glacier. *Science* **281**, 549–551 (1998).  
**As the first study to discover unstable retreat of a West Antarctic glacier in satellite data, this is a landmark paper in glaciology that has triggered widespread scientific interest in the Amundsen Sea sector.**
86. Park, J. W. et al. Sustained retreat of the Pine Island Glacier. *Geophys. Res. Lett.* **40**, 2137–2142 (2013).
87. Thoma, M., Jenkins, A., Holland, D. & Jacobs, S. Modelling Circumpolar Deep Water intrusions on the Amundsen Sea continental shelf, Antarctica. *Geophys. Res. Lett.* **35**, https://doi.org/10.1029/2008GL034939 (2008).
88. Konrad, H. et al. Uneven onset and pace of ice-dynamical imbalance in the Amundsen Sea Embayment, West Antarctica. *Geophys. Res. Lett.* **44**, 910–918 (2017).
89. Joughin, I., Alley, R. B. & Holland, D. M. Ice-sheet response to oceanic forcing. *Science* **338**, 1172–1176 (2012).  
**This review provides a great introduction to ice–ocean interactions, and how satellite observations have informed our understanding of key processes.**
90. Mercer, J. H. West Antarctic ice sheet and CO<sub>2</sub> greenhouse effect: a threat of disaster. *Nature* **271**, 321–325 (1978).
91. Anderson, J. B., Shipp, S. S., Lowe, A. L., Wellner, J. S. & Mosola, A. B. The Antarctic Ice Sheet during the Last Glacial Maximum and its subsequent retreat history: a review. *Quat. Sci. Rev.* **21**, 49–70 (2002).
92. Joughin, I., Smith, B. E. & Holland, D. M. Sensitivity of 21st century sea level to ocean-induced thinning of Pine Island Glacier, Antarctica. *Geophys. Res. Lett.* **37**, https://doi.org/10.1029/2010GL044819 (2010).
93. Joughin, I., Smith, B. E. & Medley, B. Marine ice sheet collapse potentially under way for the Thwaites Glacier basin, West Antarctica. *Science* **344**, 735–738 (2014).
94. Rignot, E., Mouginot, J., Morlighem, M., Seroussi, H. & Scheuchl, B. Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011. *Geophys. Res. Lett.* **41**, 3502–3509 (2014).
95. Milillo, P. et al. On the short-term grounding zone dynamics of Pine Island Glacier, West Antarctica, observed with COSMO-SkyMed interferometric data. *Geophys. Res. Lett.* **44**, 10436–10444 (2017).
96. Konrad, H. et al. Net retreat of Antarctic glacier grounding lines. *Nat. Geosci.* **11**, 258–262 (2018).
97. Dutrieux, P. et al. Strong sensitivity of Pine Island ice-shelf melting to climatic variability. *Science* **343**, 174–178 (2014).
98. Gloersen, P. et al. Satellite passive microwave observations and analysis of Arctic and Antarctic sea ice, 1978–1987. *Ann. Glaciol.* **17**, 149–154 (1993).
99. Zwally, H. J., Yi, D., Kwok, R. & Zhao, Y. ICESat measurements of sea ice freeboard and estimates of sea ice thickness in the Weddell Sea. *J. Geophys. Res. Oceans* **113**, https://doi.org/10.1029/2007JC004284 (2008).  
**This study was the first to attempt an extensive assessment of Antarctic sea ice thickness based on satellite altimeter measurements of floe freeboard.**
100. Kurtz, N. T. & Markus, T. Satellite observations of Antarctic sea ice thickness and volume. *J. Geophys. Res. Oceans* **117**, https://doi.org/10.1029/2012JC008141 (2012).
101. Heil, P., Fowler, C. W. & Lake, S. E. Antarctic sea-ice velocity as derived from SSM/I imagery. *Ann. Glaciol.* **44**, 361–366 (2006).
102. Holland, P. R. & Kwok, R. Wind-driven trends in Antarctic sea-ice drift. *Nat. Geosci.* **5**, 872–875 (2012).
103. Brierley, A. S. & Thomas, D. N. in *Advances in Marine Biology* Vol. **43**, 171–276 (Academic Press, 2002).
104. Massom, R. A. et al. Examining the interaction between multi-year landfast sea ice and the Mertz Glacier Tongue, East Antarctica: another factor in ice sheet stability? *J. Geophys. Res. Oceans* **115**, https://doi.org/10.1029/2009JC006083 (2010).
105. Robel, A. A. Thinning sea ice weakens buttressing force of iceberg mélange and promotes calving. *Nat. Commun.* **8**, 14596 (2017).
106. Miles, B. W. J., Stokes, C. R. & Jamieson, S. S. R. Simultaneous disintegration of outlet glaciers in Porpoise Bay (Wilkes Land), East Antarctica, driven by sea ice break-up. *Cryosphere* **11**, 427–442 (2017).
107. Turner, J., Hosking, J. S., Bracegirdle, T. J., Marshall, G. J. & Phillips, T. Recent changes in Antarctic Sea Ice. *Phil. Trans. R. Soc. A* **373**, https://doi.org/10.1098/rsta.2014.0163 (2015).
108. Armour, K. C. & Bitz, C. M. in *US Clivar Variations* Vol. **13**, 12–19 (2015).
109. Meier, W., Gallaher, D. & Campbell, G. G. New estimates of Arctic and Antarctic sea ice extent during September 1964 from recovered Nimbus I satellite imagery. *Cryosphere* **7**, 699–705 (2013).
110. Gallaher, D. W., Campbell, G. G. & Meier, W. N. Anomalous variability in Antarctic sea ice extents during the 1960s with the use of Nimbus data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**, 881–887 (2014).
111. de la Mare, W. K. Changes in Antarctic sea-ice extent from direct historical observations and whaling records. *Clim. Change* **92**, 461–493 (2009).
112. Massonnet, F., Guemas, V., Fuékar, N. S. & Doblas-Reyes, F. J. The 2014 high record of Antarctic sea ice extent. *Bull. Am. Meteorol. Soc.* **96**, S163–S167 (2015).
113. Turner, J. et al. Unprecedented springtime retreat of Antarctic sea ice in 2016. *Geophys. Res. Lett.* **44**, 6868–6875 (2017).
114. Stuecker, M. F., Bitz, C. M. & Armour, K. C. Conditions leading to the unprecedented low Antarctic sea ice extent during the 2016 austral spring season. *Geophys. Res. Lett.* **44**, 9008–9019 (2017).
115. Zhang, J. Increasing Antarctic sea ice under warming atmospheric and oceanic conditions. *J. Clim.* **20**, 2515–2529 (2007).
116. Hobbs, W. R. et al. A review of recent changes in Southern Ocean sea ice, their drivers and forcings. *Global Planet. Change* **143**, 228–250 (2016).
117. Kwok, R. & Comiso, J. C. Southern Ocean climate and sea ice anomalies associated with the Southern Oscillation. *J. Clim.* **15**, 487–501 (2002).
118. Stammerjohn, S., Massom, R., Rind, D. & Martinson, D. Regions of rapid sea ice change: an inter-hemispheric seasonal comparison. *Geophys. Res. Lett.* **39**, https://doi.org/10.1029/2012GL050874 (2012).
119. Kwok, R., Comiso, J. C., Lee, T. & Holland, P. R. Linked trends in the South Pacific sea ice edge and Southern Oscillation Index. *Geophys. Res. Lett.* **43**, 10295–10302 (2016).
120. Turner, J. & Comiso, J. Solve Antarctica's sea-ice puzzle. *Nature* **547**, 275–277 (2017).
121. Gloersen, P. Modulation of hemispheric sea-ice cover by ENSO events. *Nature* **373**, 503–506 (1995).
122. Lefebvre, W., Goosse, H., Timmermann, R. & Fichefet, T. Influence of the Southern Annular Mode on the sea ice–ocean system. *J. Geophys. Res. C* **109**, https://doi.org/10.1029/2004JC002403 (2004).
123. Holland, M. M., Landrum, L., Kostov, Y. & Marshall, J. Sensitivity of Antarctic sea ice to the Southern Annular Mode in coupled climate models. *Clim. Dyn.* **49**, 1813–1831 (2017).
124. Turner, J. et al. Non-annular atmospheric circulation change induced by stratospheric ozone depletion and its role in the recent increase of Antarctic sea ice extent. *Geophys. Res. Lett.* **36**, https://doi.org/10.1029/2009GL037524 (2009).
125. Bintanja, R., Van Oldenborgh, G. J., Drijfhout, S. S., Wouters, B. & Katsman, C. A. Important role for ocean warming and increased ice-shelf melt in Antarctic sea-ice expansion. *Nat. Geosci.* **6**, 376–379 (2013).
126. Pauling, A. G., Smith, I. J., Langhorne, P. J. & Bitz, C. M. Time-dependent freshwater input from ice shelves: impacts on Antarctic Sea Ice and the Southern Ocean in an Earth System model. *Geophys. Res. Lett.* **44**, 10454–10461 (2017).
127. Perovich, D. K. & Richter-Menge, J. A. Loss of sea ice in the Arctic. *Annu. Rev. Mar. Sci.* **1**, 417–441 (2009).
128. Comiso, J. C. & Nishio, F. Trends in the sea ice cover using enhanced and compatible AMSR-E, SSM/I, and SMMR data. *J. Geophys. Res. Oceans* **113**, https://doi.org/10.1029/2007JC004257 (2008).
129. Kwok, R. Ross Sea ice motion, area flux, and deformation. *J. Clim.* **18**, 3759–3776 (2005).
130. Hollands, T., Haid, V., Dierking, W., Timmermann, R. & Ebner, L. Sea ice motion and open water area at the Ronne Polynia, Antarctica: synthetic aperture radar observations versus model results. *J. Geophys. Res. Oceans* **118**, 1940–1954 (2013).

131. Emery, W. J., Fowler, C. W. & Maslanik, J. A. Satellite-derived maps of Arctic and Antarctic sea ice motion: 1988 to 1994. *Geophys. Res. Lett.* **24**, 897–900 (1997).  
**This paper is an early application of repeat satellite imagery for tracking the motion on sea ice floes in the polar regions, demonstrating that the Southern Hemisphere sea ice pack tends to drifts northwards under the influence of ocean currents and katabatic winds.**
132. Kwok, R., Pang, S. S. & Kacimi, S. Sea ice drift in the Southern Ocean: regional patterns, variability, and trends. *Elem. Sci. Anth.* **5**, <https://doi.org/10.1525/elementa.226> (2017).
133. Alley, K. E., Scambos, T. A., Siegfried, M. R. & Fricker, H. A. Impacts of warm water on Antarctic ice shelf stability through basal channel formation. *Nat. Geosci.* **9**, 290–293 (2016).
134. Giles, K. A., Laxon, S. W. & Worby, A. P. Antarctic sea ice elevation from satellite radar altimetry. *Geophys. Res. Lett.* **35**, <https://doi.org/10.1029/2007GL031572> (2008).
135. Willatt, R. C., Giles, K. A., Laxon, S. W., Stone-Drake, L. & Worby, A. P. Field investigations of Ku-band radar penetration into snow cover on Antarctic sea ice. *IEEE Trans. Geosci. Remote Sens.* **48**, 365–372 (2010).
136. Tin, T. & Jeffries, M. O. Sea-ice thickness and roughness in the Ross Sea, Antarctica. *Ann. Glaciol.* **33**, 187–193 (2001).
137. Farrell, S. L. et al. Sea-ice freeboard retrieval using digital photon-counting laser altimetry. *Ann. Glaciol.* **56**, 167–174 (2015).
138. Markus, T. et al. The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): science requirements, concept, and implementation. *Remote Sens. Environ.* **190**, 260–273 (2017).
139. Armitage, T. W. K. & Ridout, A. L. Arctic sea ice freeboard from AltiKa and comparison with CryoSat-2 and Operation IceBridge. *Geophys. Res. Lett.* **42**, 6724–6731 (2015).
140. Guerreiro, K., Fleury, S., Zakharova, E., Rémy, F. & Kouraev, A. Potential for estimation of snow depth on Arctic sea ice from CryoSat-2 and SARAL/AltiKa missions. *Remote Sens. Environ.* **186**, 339–349 (2016).
141. Fricker, H. A. & Padman, L. Ice shelf grounding zone structure from ICESat laser altimetry. *Geophys. Res. Lett.* **33**, <https://doi.org/10.1029/2006GL026907> (2006).
142. Haran, T., Bohlander, J., Scambos, T., Painter, T. & Fahnestock, M. A. *MODIS Mosaic of Antarctica 2008–2009 (MOA2009) Image Map*. (National Snow and Ice Data Center (NSIDC), Boulder, 2014).
143. Tschudi, M., Fowler, C. W., Maslanik, J. A., Stewart, J. S. & Meier, W. *EASE-Grid Sea Ice Age*. Version 3. (NASA NSIDC Distributed Active Archive Center, Boulder, 2016).
144. Fetterer, F., Knowles, K., Meier, W., Savoie, M. & Windnagel, A. K. *Sea Ice Index*. Version 2 (NSIDC, Boulder, Colorado USA, 2017).
145. Ryan, W. B. F. et al. Global multi-resolution topography synthesis. *Geochem. Geophys. Geosyst.* **10**, <https://doi.org/10.1029/2008GC002332> (2009).
146. Timmermann, R. et al. A consistent dataset of Antarctic ice sheet topography, cavity geometry, and global bathymetry. *Earth Syst. Sci. Data* **2**, 261–273, <https://doi.org/10.1594/pangea.741917> (2010).
147. Locarnini, R. A. et al. *World Ocean Atlas 2009 Vol. 1 Temperature*. NOAA Atlas NESDIS 68 (eds Levitus, S.) <http://www.nodc.noaa.gov/OC5/indprod.html> (US Government Printing Office, Washington DC, 2010).
148. McMillan, M. et al. Increased ice losses from Antarctica detected by CryoSat-2. *Geophys. Res. Lett.* **41**, 3899–3905 (2014).
149. Anandakrishnan, S. & Alley, R. B. Stagnation of ice stream C, West Antarctica by water piracy. *Geophys. Res. Lett.* **24**, 265–268 (1997).

**Acknowledgements** This work was supported by the UK Natural Environment Research Council's Centre for Polar Observation and Modelling (cpom300001) and the European Space Agency's Climate Change Initiative. AS was supported by a Royal Society Wolfson Research Merit award. SLF was supported under NASA grant 80NSSC17K0006 and NOAA grant NA14NES4320003. We thank T. Slater, A. Ridout, and L. Gilbert for their help in preparing Fig. 1 and Fig. 2, and K. Duncan for help in preparing Fig. 4.

**Author contributions** A.S. coordinated the work, and led the review of grounded ice. H.F. led the review of ice shelves and subglacial lakes. S.F. led the review of sea ice.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to A.S.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Choosing the future of Antarctica

S. R. Rintoul<sup>1,2,3\*</sup>, S. L. Chown<sup>4</sup>, R. M. DeConto<sup>5</sup>, M. H. England<sup>6</sup>, H. A. Fricker<sup>7</sup>, V. Masson-Delmotte<sup>8</sup>, T. R. Naish<sup>9</sup>, M. J. Siegert<sup>10</sup> & J. C. Xavier<sup>11,12</sup>

**We present two narratives on the future of Antarctica and the Southern Ocean, from the perspective of an observer looking back from 2070. In the first scenario, greenhouse gas emissions remained unchecked, the climate continued to warm, and the policy response was ineffective; this had large ramifications in Antarctica and the Southern Ocean, with worldwide impacts. In the second scenario, ambitious action was taken to limit greenhouse gas emissions and to establish policies that reduced anthropogenic pressure on the environment, slowing the rate of change in Antarctica. Choices made in the next decade will determine what trajectory is realized.**

**A**ntarctica, the most remote region on Earth, is intimately coupled to the rest of the climate system. Atmospheric and oceanic teleconnections communicate climate variations at low latitude to Antarctica and the Southern Ocean, influencing the polar atmosphere, ocean, ice sheet, sea ice and biosphere. Likewise, Antarctica and the surrounding Southern Ocean affect the rest of the globe. The amount and rate of sea level rise in the coming centuries depends on the response of the Antarctic Ice Sheet to warming of the atmosphere and ocean<sup>1</sup>. The Southern Ocean takes up more anthropogenic heat and carbon than the oceans at other latitudes, helping to slow the pace of atmospheric warming<sup>2,3</sup>. The circulation of the Southern Ocean also sustains global marine productivity by returning nutrient-rich deep water to the surface and exporting nutrients to lower latitudes<sup>4</sup>. Given the profound influence of Antarctica and the Southern Ocean on sea level, climate, and marine ecosystems, change in the region will have widespread consequences for the Earth and humanity. From a political perspective, Antarctica and the Southern Ocean are among the largest shared spaces on Earth, regulated by the unique governance regime of the Antarctic Treaty System<sup>5</sup>, and embedded within and connected to broader global decision-making<sup>6,7</sup>.

We present, from the perspective of an observer in 2070, two narratives on 50 years of change in Antarctica. Each narrative highlights the long-term consequences of decisions made today. The 50-year timescale reflects a period over which substantial differences between the two scenarios will develop and is within the lifetime of today's children. In the first, no meaningful action was taken to mitigate greenhouse gas emissions and global warming continued unabated. In the second, aggressive measures were taken to limit emissions, restrict global warming and increase resilience. We consider the 'high emissions/weak action' and 'low emissions/strong action' narratives to be likely upper and lower bounds on the future trajectory of Antarctica and the Southern Ocean. The trajectory that plays out over the next 50 years depends on choices made today. Cumulative emissions of CO<sub>2</sub> largely determine global mean surface warming<sup>1</sup>, so continued growth in emissions soon commits us to further unavoidable climate impacts, even if some of those impacts take decades or centuries to emerge fully<sup>8</sup>. Greenhouse gas emissions must start decreasing in the coming decade to have a realistic prospect of following the low emissions narrative<sup>9</sup>.

We provide an integrated assessment of the associated trajectories for Antarctica and the Southern Ocean, spanning physical, biological

and social dimensions. These trajectories describe plausible alternative futures rather than forecasts. Where possible, such as for some physical and chemical variables, we use quantitative projections from climate and ice sheet models. Where this is not possible (as for many biological and social systems), we anticipate future change using a heuristic approach based on process understanding and the known response to past changes (noting that the heuristic approach risks neglecting important nonlinearities or surprises in natural systems<sup>10,11</sup>).

Of necessity, we offer the perspective of a single observer on 50 years of change in Antarctica, but we acknowledge that given the global diversity of human experience and values, other observers would interpret these changes through a different lens. We do not assume that individuals, institutions or society will act in a certain way; rather, we assess the possible consequences of particular courses of action (or inaction). Our goal is to initiate discussion and consideration of options for Antarctica's future, and to highlight how the future of Antarctica is tied to that of the rest of the planet and human society.

## Antarctica in 2070 under high emissions

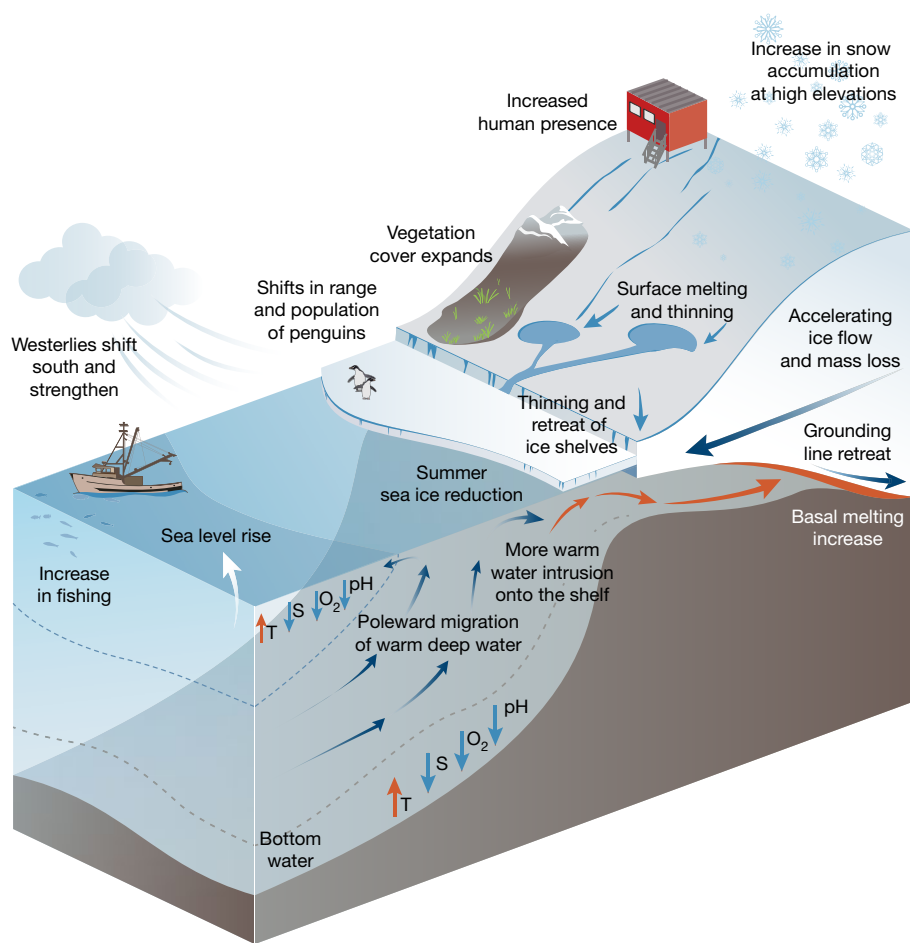
Looking back from 2070, it is clear that the past 50 years have unfolded much as anticipated by the high emissions (Representative Concentration Pathway (RCP) 8.5) scenario used by the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report<sup>1</sup> published more than half a century ago. Growth in energy and food demand by the increased global population, supplied predominantly by intensive agriculture supported by fossil fuel use and associated with deforestation, drove an ongoing acceleration of greenhouse gas concentrations in the atmosphere and an increase in effective radiative forcing of about 5 W m<sup>-2</sup> compared to the pre-industrial period<sup>1</sup>. Lack of action to mitigate greenhouse gas emissions was accompanied by lack of regulation of the human presence in Antarctica. Both distant and local human activities have left an indelible footprint on the Antarctic and Southern Ocean environment (Fig. 1).

## Change in the physical environment

After 50 years of continued high greenhouse gas emissions, global mean surface air temperatures over land are now more than 3.5 °C higher than observed in the late nineteenth century<sup>1</sup> (Fig. 2), well above the symbolic 'guardrail' of 2 °C introduced in international climate agreements

<sup>1</sup>CSIRO Oceans & Atmosphere, Hobart, Tasmania, Australia. <sup>2</sup>Antarctic Climate and Ecosystems Cooperative Research Centre, Hobart, Tasmania, Australia. <sup>3</sup>Centre for Southern Hemisphere Oceans Research, Hobart, Tasmania, Australia. <sup>4</sup>School of Biological Sciences, Monash University, Victoria, Australia. <sup>5</sup>University of Massachusetts, Amherst, MA, USA. <sup>6</sup>ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales, Australia. <sup>7</sup>Scripps Institution of Oceanography, La Jolla, CA, USA. <sup>8</sup>LSCE (IPSL, CEA-CNRS-UVSQ, Université Paris Saclay), Paris, France. <sup>9</sup>Victoria University of Wellington, Wellington, New Zealand. <sup>10</sup>Grantham Institute and Department of Earth Science and Engineering, Imperial College London, London, UK. <sup>11</sup>Marine and Environmental Science Centre MARE, Department of Life Sciences, University of Coimbra, Coimbra, Portugal. <sup>12</sup>British Antarctic Survey, Natural Environment Research Council, Cambridge, UK. \*e-mail: Steve.Rintoul@csiro.au





**Fig. 1 | Schematic summary of impacts on Antarctica and the Southern Ocean in 2070, under a ‘high emissions/low action’ scenario.** T, temperature; S, salinity; O<sub>2</sub>, oxygen.

after the 2009 United Nations Framework Convention on Climate Change meeting in Copenhagen and reaffirmed by the nations of the world in a landmark agreement signed in Paris in December 2015.

**Atmosphere.** Air temperatures over Antarctica have warmed<sup>1</sup> by about 3 °C, well above the range of centennial variations of the current interglacial period<sup>12</sup>. In some low-lying regions of the continent, this increase in temperature has been sufficient to cause surface melt in summer. The increase in summer melt has had widespread impacts, contributing to the collapse of ice shelves, exposing new ice-free areas open to colonization by native and introduced plants, and altering the area suitable for nesting penguins and other birds.

During the 30 years between 2015 and 2045, recovery of the ozone hole tended to push the westerlies further north, while increases in greenhouse gas concentrations caused a shift to the south. As a result, spring and summer wind patterns changed little over this time period, in stark contrast to the strengthening and poleward shift of the westerlies seen in summer in previous decades<sup>13</sup>. During the winter season, when the ozone hole had little impact on the winds, the westerlies began to shift polewards and intensify after 2010 in response to increasing greenhouse gas concentrations in the atmosphere. By around 2045, greenhouse gases had won out over ozone recovery even in summer, forcing the westerlies to shift south and strengthen in all seasons<sup>14</sup>.

**Southern Ocean.** Southern Ocean surface waters have warmed everywhere, reversing a slight cooling observed in the early twenty-first century at high latitudes, with an average increase<sup>1</sup> of 1.9 °C south of 50 °S. Surface waters have also freshened in response to increased precipitation and melt of sea ice and glacier ice<sup>15–17</sup>. Warming and freshening of surface waters have increased stratification and inhibited exchange with nutrient-rich deep waters. At 1,000 m depth, warming north of the

Antarctic Circumpolar Current has exceeded ocean temperature rise observed at other latitudes, reflecting the efficient uptake of anthropogenic heat by the overturning circulation<sup>3</sup>. Antarctic Bottom Water, the volume of which had already contracted by 50% between 1970 and 2014<sup>18</sup>, no longer exists. As a result of freshening, water sinking near Antarctica is no longer dense enough to qualify as Antarctic Bottom Water, although continued sinking of less-dense water maintains deep oxygen levels<sup>19</sup>. In response to ongoing changes in surface winds, the subtropical gyres have shifted polewards, effectively shrinking the Southern Ocean<sup>20</sup>.

The combined impact of changes in winds, warming and freshening has slightly strengthened the upper limb of the Southern Ocean overturning circulation<sup>21</sup>. The Southern Ocean therefore continues to take up and export large amounts of heat and carbon dioxide, helping to slow the increase in global surface temperatures. Although uptake and export of heat by the overturning circulation initially delayed warming around Antarctica, surface sea and air temperatures in and around Antarctica are now warming at about the same rate as the global average, with larger warming in winter than in summer<sup>1</sup>.

The chemistry of Southern Ocean waters continued to change in response to rising levels of atmospheric CO<sub>2</sub>. The pH of surface waters south of 60° S decreased by 0.2 between 2017 and 2070, equivalent to a 50% decrease in the concentration of hydrogen ions since the pre-industrial period<sup>1</sup>. Southern Ocean surface waters south of 60° S became under-saturated with respect to aragonite in winter by 2040, and by 2070, >30% of the Southern Ocean surface waters south of 40° S had become under-saturated year-round<sup>1,22</sup>. Hence, waters have become corrosive to shells and other biological structures made of this form of calcium carbonate.

**Ice shelves.** Wind-driven changes in ocean currents resulted in an increase in ocean heat transport to the Amundsen Sea in the late

twentieth and early twenty-first century<sup>23</sup>. Warmer ocean waters entering the cavities beneath floating ice shelves drove higher rates of basal melting, thinning of ice shelves, and a reduction in the back-stress on the grounded ice upstream. The reduced buttressing increased the flow of the ice streams feeding the ice shelves and led to the retreat of grounding lines, including runaway retreat of glaciers grounded on bedrock that deepened inland<sup>24–27</sup>.

Although some ice shelves had supported extensive surface melting for decades<sup>28</sup>, summer air temperatures are now high enough to increase surface melt on large areas of the floating ice shelves<sup>29–31</sup>. The increased volume of surface melt, coupled with an increase in the temperature of the surface firn due to persistent refreezing of meltwater and associated release of latent heat, has now led to the collapse of several ice shelves through hydrofracturing, a process first observed in Antarctica 70 years ago during the 2002 collapse of the Larsen B Ice Shelf<sup>32</sup>. The Larsen C Ice Shelf collapsed by hydrofracture in 2015<sup>33</sup> following several decades of thinning<sup>34</sup>, and then several consecutive summers of excessive summer melting.

Most of the ice shelves in the Amundsen Sea have thinned at an accelerating rate owing to increased ocean temperatures that caused higher basal melt rates in the sub-ice cavities. The Venable, Crosson and Dotson ice shelves were all lost between 2040 and 2050, quickly followed by the Thwaites ice shelf in 2060, as anticipated in 2015 from trends measured by satellites between 1994 and 2012<sup>34</sup>. Nearly a quarter of the volume of Antarctica's ice shelves has been lost in the past 50 years<sup>33</sup>. Loss of sea ice also contributed to a decrease in buttressing of ice shelves<sup>35</sup>. Totten Glacier, an outlet for a large ice-sheet drainage basin in East Antarctica, has undergone thinning and retreat driven by warm water<sup>36</sup> accessing the grounding line. Changes in ocean currents have led to warmer ocean water entering ice shelf cavities (as predicted for the Filchner Ronne Ice Shelf<sup>37</sup>), causing ice-shelf thinning and the retreat of the grounding line across ice streams identified previously as being particularly vulnerable to such change in both the West<sup>38</sup> and East Antarctic<sup>39</sup> ice sheets.

The large number of icebergs produced by collapsing ice shelves all around Antarctica<sup>40</sup> is now carefully monitored to manage the risks to the greatly expanded fishing, tourism and commercial shipping fleets, and Antarctic operations by Antarctic Treaty nations.

**Antarctic Ice Sheet and sea level.** Fifty years ago, mass loss from the ice-sheet margins of West Antarctica was partially compensated by mass gain due to increased snowfall over East Antarctica<sup>41,42</sup>, facilitated by more frequent intrusions of marine air as the westerlies shifted south<sup>43</sup>. However, the increase in ocean-driven melting could not be balanced by enhanced accumulation after 2020, leading to unequivocal loss of mass from the Antarctic Ice Sheet.

Observations of grounding line locations and ice stream velocities now confirm that the 'marine ice sheet instability' is well underway in West Antarctica, as first proposed in the 1970s<sup>24</sup> and supported by observations and modelling in 2014<sup>25–27</sup>. The loss of back-stress after the disappearance of ice shelves has led to increased flow of ice from the ice sheet to the ocean in both East and West Antarctica. Tall, unstable<sup>44</sup> ice cliffs have begun to appear in places around the marine-terminating ice sheet margin, where ice shelves and glacier tongues that were more than 750 m thick at their grounding zones have been lost. Further collapse of the West Antarctic Ice Sheet is now irreversible<sup>26,27</sup>, mainly through the rapidly retreating, 120-km-wide Thwaites Glacier. Mass loss from the Antarctic Ice Sheet has contributed more than 27 cm of global sea level rise since 2000, as predicted decades earlier<sup>33,45</sup>. The rate of mass loss from Antarctica now exceeds 5 mm yr<sup>-1</sup> (in sea level equivalent terms) and continues to accelerate. Antarctica now makes the largest contribution to the rise in global mean sea level, exceeding the contribution from thermal expansion, the retreat of mountain glaciers and melting of the Greenland Ice Sheet. The total rate of sea level rise is similar to rates during the last deglaciation (averaging 10–15 mm yr<sup>-1</sup>)<sup>1</sup>. A commitment to multiple metres of sea level rise in the longer term is now irreversible,

consistent with early ice sheet model projections<sup>33,45–47</sup> and sea-level reconstructions of past warmer worlds<sup>48</sup>, including the Pliocene (3 million years ago, when atmospheric CO<sub>2</sub> concentrations were only 400 parts per million by volume), and the most recent interglacial period (125 thousand years ago). High emissions over the past 50 years have already committed us to more than 10 m of sea level rise in the longer term (that is, >3,000 years); if emissions continue on the present trajectory to a total of 5,120 Pg of carbon, we are committed to more than 50 m of sea level rise in 10,000 years, 80% of which will be contributed by the Antarctic Ice Sheet<sup>49</sup>. Economic losses from the flooding of coastal cities already exceed US\$1 trillion per year<sup>50</sup> as a result of the sea level rise of less than half a metre that occurred between 2000 and 2070<sup>1,33</sup>.

## Sea ice

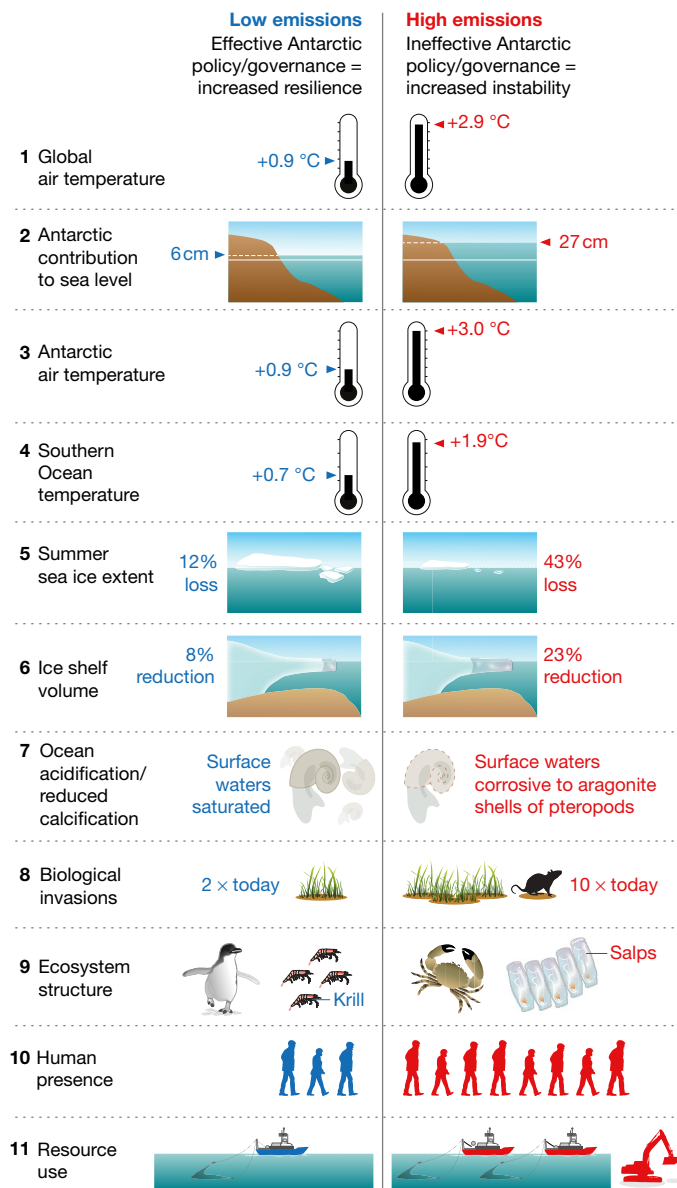
Slow expansion of Antarctic sea ice in the decades preceding 2016 encouraged, in some, a sense of complacency about the stability of Antarctic sea ice. Record low sea ice extent in 2016 served as a cautionary reminder that this past stability could not be taken for granted. Indeed, by 2045 Antarctic sea ice was in clear and sustained retreat. Winter sea ice extent has reduced by 40%, more than twice the retreat projected for 2070 by climate models<sup>1</sup>, consistent with the large sensitivity of winter sea ice extent to Antarctic warming inferred from palaeoclimate information<sup>51</sup>. Summer sea ice extent has decreased by almost half and most of the continental margin is now regularly free of sea ice in February, enabling access by ship to previously inaccessible regions, increasing fishing activity and tourism.

## Change in biology and ecosystems

**Fisheries.** As access to the Antarctic coast has become easier, Antarctic marine systems are being exploited by fisheries from many nations, with stocks in decline around the continent. While initial conservation actions, including new marine protected areas, provided some respite for toothfish (*Dissostichus eleginoides* and *Dissostichus mawsoni*)<sup>7</sup>, their long life cycles<sup>52</sup>, illegal and unregulated practices<sup>53</sup>, and increased interest in the species because of fishery collapses elsewhere<sup>54</sup> quickly depleted their populations. Interest in the Antarctic krill *Euphausia superba* fishery has seen profound growth because of new technologies for product processing, and resource depletion elsewhere<sup>55</sup>. Coupled with growing impacts of acidification on juvenile stages<sup>56</sup>, replacement of Antarctic krill by salps in many areas owing to sea ice changes<sup>57</sup>, and a rise in the number of baleen whales<sup>58</sup> the fishery has substantially reduced stocks, at least as far as assessments show, though these have been compromised by the absence of an extensive research-based assessment<sup>59</sup>. In the ensuing scramble to make use of resources, the evidence-based management approaches implemented successfully by the Commission on the Conservation of Antarctic Marine Living resources (CCAMLR) were overwhelmed<sup>58,60</sup>.

**Predators.** The decline of exploited fish and krill stocks affected the populations of predators, including declines of several populations of penguin species<sup>61,62</sup>, though with complexity in some areas arising from declines in competition with exploited toothfish<sup>63</sup>. Some Antarctic krill predators, such as black-browed and grey-headed albatrosses, have shifted their diet from Antarctic krill to mesopelagic fish and squid in response<sup>64</sup>, joining a large suite of myctophid feeders<sup>65</sup>, including fur seals<sup>66</sup>. Fishery-related mortality of seabirds and seals has increased because of growth in mesopelagic fish and squid fisheries<sup>67–69</sup>. Overall, Southern Ocean fisheries are declining in value and their regulation is increasingly contentious, reflecting problems encountered globally in fisheries during the twentieth and twenty-first centuries.

**Community structure.** Warming has also led to substantial changes in the composition of marine communities. In the decades following 2017, regional trends of increase in some species and declines in others continued<sup>70,71</sup>. More recently, unexpected regime shifts occurred,



**Fig. 2 | Antarctica and the Southern Ocean in 2070, under ‘low emissions/high action’ (left) and ‘high emissions/low action’ (right) scenarios.** Differences are relative to a 1986–2005 reference period. Differences (1) to (6) are atmosphere–ocean–ice differences, taken from model projections following low- and high-emissions scenarios, respectively. Data for differences (1), (3), (4) and (5) are from ref. <sup>1</sup>. Data for differences (2) and (6) are from ref. <sup>33</sup>. Differences (7) to (11) are differences in ecosystem states and pressures, with ecosystem structure changing from the current situation to one characterized by new species and interactions. Data for difference (7) is from refs <sup>1</sup> and <sup>22</sup>; for (8) from refs <sup>82</sup> and <sup>83</sup>; for (9) from refs <sup>57,78</sup> and <sup>113</sup>; for (10) from ref. <sup>6</sup> and for (11) from ref. <sup>5</sup>. The low-emissions scenario sees greenhouse gas mitigation adhered to, limiting global warming by 2070 to 0.9 °C above the 1986–2005 mean. The high-emissions scenario, in which no mitigation takes place, leads to 2.9 °C of global warming by 2070 relative to the 1986–2005 mean, or 3.5 °C relative to 1850–1900. The systems assessed are: (1) global average air temperature; (2) Antarctic contribution to global sea level; (3) Antarctic surface air temperature; (4) Southern Ocean surface temperature; (5) summer (February) sea ice extent; (6) Antarctic ice shelf volume; (7) ocean acidification (illustrated by a pteropod, a marine snail, with an aragonite shell subject to dissolution under acidic conditions); (8) level of alien species invasion; (9) ecosystem structure (under the low-emissions scenario the present ecosystem continues; under the high-emissions scenario some species, such as crabs, become established and other species shifts occur, such as from krill to salps, as the climate warms and sea ice retreats); (10) human presence; and (11) resource use. Each of these systems will continue to change after 2070, with the magnitude of the change to which we are committed being generally much larger than the change realized by 2070.

caused by changing interactions among key species (such as between Antarctic krill, penguins, seals and baleen whales), catastrophic declines in some species, and in response to new phenomena such as transport of soil particles to the ocean by increased run-off of ice melt from the continent<sup>72,73</sup>. Changes in resource availability further hastened Adelie and Chinstrap penguin range contractions<sup>61,74</sup>. Gentoo penguins benefited initially, as early assessments predicted<sup>75</sup>, but warming to the north and fishery impacts have led to trailing-edge range contractions and the start of population declines in this species too. Ocean acidification further complicated community responses. While some species adapted to acidification (for example, Antarctic brachiopods<sup>76</sup>), others were less able to do so (for example, pteropods<sup>77</sup>), resulting in reorganization of communities<sup>70</sup>, compounded by changing ocean and atmospheric conditions<sup>78</sup>.

**Terrestrial biota and invasive species.** On land, melt and retreat of glaciers exposed new ice-free areas, particularly on the Antarctic Peninsula, the northernmost part of the continent<sup>79</sup>. Antarctica’s only two native vascular plant species showed initial increases in populations and expanded their ranges widely on the peninsula, with Antarctic hair grass predominating<sup>80</sup>. By 2050 this vegetation had come to include many other species too. Some are recognizable as widespread Antarctic invaders, especially the annual bluegrass or *Poa annua*,

already recorded from the peninsula in the early part of the century<sup>81</sup>. Others are species that had long been predicted to colonize on the basis of analyses of seeds carried to Antarctica by scientific and tourist operations<sup>82</sup>. For other species, provenance remains obscure: they could have colonized naturally or through human agency. Debate of this contentious point at the Antarctic Treaty’s Committee for Environmental Protection (CEP) precluded action, resulting in their further spread. By 2070, many research stations and several new tourist hotels have, in consequence, developed manicured gardens. Settlements of permanent residents, including small numbers of migrants, have now become established to service the research and tourism industries<sup>6</sup>, and to control invasive pests<sup>83</sup>. Invasive species management lessons learned in the late twentieth century from the sub-Antarctic islands seem to have been forgotten.

Elsewhere on the continent, non-indigenous species have yet to establish populations<sup>82,83</sup>. Yet, owing to discord in the CEP precipitated by rapidly changing conditions on the Antarctic Peninsula and many biotic invasions, attention to transfer of species among the continent’s conservation biogeographic regions diminished<sup>84,85</sup>. Molecular phylogeographic studies of several groups, and especially the microbiota, were discontinued after investigations revealed that exchange of species and populations among Antarctic biogeographic regions had become virtually continuous after 2050<sup>86</sup>.



**Biodiversity conservation.** Both in the Southern Ocean and on continental Antarctica, action early in the twenty-first century led to improvements in biodiversity conservation, especially after attention was drawn to the potential for effective management to deliver rapid improvements in Antarctic biodiversity<sup>7</sup>. By 2050, however, many gains had started to suffer attrition. Establishment of protected areas slowed after an increase in the first two decades of the century, and protected area management failed to keep up with threats from growing human activity and resource exploitation<sup>60,87</sup>. Moreover, funding of environmental protection to underpin decision-making declined owing to larger environmental degradation problems elsewhere, including those associated with rapidly rising sea levels<sup>88</sup>. Although the CEP continues to meet on an annual basis, its recommendations are generally ignored at meetings of the parties to the Antarctic Treaty, unless they are relevant to the dominant issues of resource apportionment and management of expanding ice-free areas and widespread, land-based tourism. The special meeting held 50 years ago at the 40<sup>th</sup> Antarctic Treaty Consultative Meeting entitled 'Our Antarctica: Protection and Utilization' presaged this change.

### Change in human engagement with Antarctica

Perceptions of the priorities for and efficacy of the Antarctic governance regime and its global role varied widely among governments, NGOs and diverse members of civil society, and changed over time in concert with a rapidly changing global order. While the reaction of different actors to the growing evidence of dramatic change in the Antarctic environment therefore diversified, overall the past 50 years have been characterized by gradual erosion of the systems that safeguarded the Antarctic environment in the late twentieth and early twenty-first century<sup>5</sup>.

**Harvesting.** Between 2017 and 2070, the global human population grew by >40%, from 7 billion to 10 billion<sup>89</sup>. Owing to tremendous pressure for resources to support this larger and more prosperous population, including a now overweight or obese majority<sup>90</sup>, Antarctica and the Southern Ocean are now even more widely explored or exploited. Most exploitation has occurred in the Southern Ocean, with a diversity of marine species harvested. At the CCAMLR, discussions now focus on resource apportionment among nations, rather than on an ecosystem-based approach to conservation<sup>60,91</sup>. On the continent, harvesting is less noticeable. Ongoing legal battles over bioprospecting<sup>7</sup> continue, however. Much of the pharmaceutical prospecting interest lies in discovering products that may manipulate human metabolism.

**Mining.** In 2049, several nations attempted to rescind Article 7 of the Protocol on Environmental Protection to the Antarctic Treaty, which prohibits mineral resource exploitation. A majority of nations agreed, but the motion failed to obtain agreement of three-quarters of the states that were Antarctic Treaty Consultative Parties at the time of the adoption of the Environmental Protocol, a requirement of Article 25. Thus, the Environmental Protocol remained unchanged. The early, clear division between the parties espousing a research for use approach and those more concerned with conservation<sup>5</sup> largely disappeared after 2048. Many nations started investigating resource potential and extraction technologies, thinly veiled under the guise of scientific exploration. Much of this 'research' activity now verges on extraction, with little political will to challenge these actions because of likely interference with geo-political complexities elsewhere on the planet<sup>6</sup>. Rapid technological developments in the Arctic have made resource exploration more affordable in polar regions, and global shortages of key minerals have driven investment in what has now come to be called 'scientific exploration'.

**Tourism.** Although a visit to Antarctica still remains the privilege of a limited few compared with the global population, tourist numbers now exceed one million each year, reflecting rapid growth after an initial hiatus in the early twenty-first century<sup>92</sup>. For many nations, tourism

revenue now provides the main source of funding for national Antarctic programmes, alongside partnerships with the fishing, pharmaceutical, food and minerals industries. Management of the Antarctic environment is now similar to that of national parks elsewhere, where managers strike an uneasy balance between revenue generation, tourist numbers and biodiversity protection. Most of the global population now lives in cities and is less connected with the importance and meaning of the natural world (an extinction of experience<sup>93</sup>), resulting in loss of commitment to environmental concerns, including Antarctica. In consequence, programmes focused on a distant continent rather than on immediate surroundings attract little media attention, apart from the wide concern about Antarctica's impact on global sea level rise. (A rearguard action to delay sea level rise by pumping seawater onto the Antarctic continent to be stored as ice, with power supplied by 850,000 1.5-MW wind turbines, fell well short of the scale needed to make much difference to sea level rise<sup>94</sup>). While the Antarctic Treaty System's various agreements remain in place, they have weakened, leaving them vulnerable to other actors on a stage marked by regional rivalry rather than international cooperation.

### Antarctica in 2070 under low emissions

Although the prospects for effective global action to mitigate emissions looked grim in 2015, the subsequent ratification of the United Nations Paris climate agreement by 196 countries, including the USA after some delay, heralded a new era of international cooperation to reduce greenhouse gas emissions. The faster-than-anticipated decrease in renewable energy and battery costs triggered a rapid transition out of coal. An increase in the magnitude and frequency of extreme climate events affecting major populations and economies highlighted widespread vulnerability and convinced decision-makers to increase their ambition to reduce greenhouse gas emissions, with the strong involvement of cities, regions and business. As a result of these policies, amplifying carbon feedbacks were not triggered, and we are now on track to keep warming well below the 2 °C target. New financial pathways helped create a functional and equitable carbon market, which incentivized business to transition rapidly to a low-carbon economy. Business leaders and fund managers began to appreciate the financial opportunities and other co-benefits of the transition associated with de-carbonization, and new technologies allowed for safe and efficient sequestration and ultimately removal of greenhouse gases from the atmosphere. As a result, radiative forcing has more or less followed the so-called RCP2.6 scenario considered by the IPCC Fifth Assessment Report, with radiative forcing reaching a peak in about 2040 and with net fossil fuel emissions now negative<sup>1</sup>. Widespread recognition of the dangers of unrestricted use of fossil fuels inspired changes in consumption patterns in the developed world, including shifts to more sustainable plant-based diets and changes in agriculture and land-use practices. The availability of low-cost renewable energy enabled developing countries to provide affordable clean energy and alleviate poverty. Progress in meeting the challenge of climate change was accompanied by a renewed global commitment to the Sustainable Development Goals, most of which were achieved by 2035<sup>95</sup>. Early action to reduce emissions allowed some costly adaptation measures to be avoided (for example, the US\$50 billion per year needed to protect coastal cities against flood losses<sup>50</sup>), freeing up funds for social goods such as improved healthcare and poverty reduction.

### Change in the physical environment

The physical environment of Antarctica and the Southern Ocean remains similar in many respects to that of 50 years ago<sup>1</sup> (Fig. 2). Climate variability in Antarctica and the Southern Ocean continues to be dominated by the Southern Annular Mode, the high-latitude atmospheric response to ENSO<sup>96</sup>, and interactions between the two<sup>97</sup>.

**Atmosphere.** Atmospheric trends observed in the decades before 2017 were largely associated with changes in the Southern Annular Mode related to the ozone hole, in particular a southward shift and strengthening of the westerly winds over the Southern Ocean, particularly in

summer<sup>13</sup>. Decreases in emissions of ozone-depleting substances as a result of the 1989 Montreal Protocol led to gradual repair of the ozone hole and ozone levels in the Antarctic stratosphere have now returned to the values of the 1960s<sup>98</sup>. Repair of the ozone hole and a stabilization of greenhouse gas concentrations in the atmosphere were accompanied by a gradual shift towards the Equator and weakening of the westerly winds in summer<sup>99</sup>, returning to values typical of the twentieth century<sup>14</sup>. Surface air and sea temperatures warmed by less than 1 °C and precipitation slightly increased (<10%) over the ocean and interior of the Antarctic continent<sup>1</sup>.

**Southern Ocean and sea ice.** Trends observed in temperature, salinity and circulation of the Southern Ocean in the late twentieth century and early twenty-first century slowed and ultimately reversed in the decades between 2020 and 2050. The return of the westerly winds to a position closer to the Equator was associated with a similar shift of the Antarctic Circumpolar Current and hence cooling in parts of the Southern Ocean. The overturning circulation continued to transfer anthropogenic heat and carbon dioxide effectively into the ocean interior. Changes in wind-driven ocean currents reduced the exposure of the floating ice shelves to basal melt by warm ocean waters. However, the reduction in ocean heat transport to the ice shelf cavities came too late to save some West Antarctic ice shelves and ice tongues. Sea ice extent declined slightly (<15%) in both summer and winter between 2015 and 2070<sup>1</sup>.

Effective action to mitigate emissions has also slowed the rate of increase in acidity of the oceans. The pH of Southern Ocean surface waters stabilized in the 2040s at values about 0.15 below pre-industrial values, or 0.05 below values observed in 2015<sup>1</sup>. Southern Ocean surface waters remain super-saturated with respect to aragonite. The exposure of Southern Ocean biota to ocean acidification has therefore increased only marginally over the past 50 years.

**Ice shelves.** While some ice shelves in the Antarctic Peninsula and Amundsen Sea were lost, the thinning rates observed in the large ice shelves for the period 1994–2012 remained fairly steady through to 2070. The Totten, Amery and Larsen C ice shelves remain largely intact, undergoing normal retreat through several large iceberg calving events. The marine ice cliff instability<sup>33,44</sup>, which glaciologists feared could become widespread by 2050, has mostly been limited to a few outlet glaciers in the Amundsen Sea sector of West Antarctica and has not reached East Antarctica. Persistence of pore space in the surface layer allowed more meltwater to be stored within the firn, decreasing the susceptibility to hydrofracture<sup>100</sup>.

**Antarctic Ice Sheet and sea level.** Although dynamic ice loss from marine-based sectors of the ice sheet has occurred, the rate of change is much less than the worst-case projections because many ice shelves continue to provide back-stress on the grounded ice<sup>33,45</sup>. Mass loss from the Amundsen Sea sector of West Antarctica continued, contributing 6 cm of sea level rise between 2000 and 2070<sup>33</sup>. After retreating steadily until 2050, the grounding zone of the Thwaites Glacier re-stabilized on a topographic feature about 25 km landwards of its early twenty-first-century position<sup>101</sup>, saving the West Antarctic Ice Sheet from further decay. In 2070, sea level rise continues to be dominated by contributions from ocean thermal expansion, glacier melt, and equal roles of the Greenland and west Antarctic ice sheets, as in the 2010s.

## Change in biology and ecosystems

**Value of Antarctica.** Following the ratification of the United Nations Paris Climate Agreement, and the Santiago Declaration by the parties to the Antarctic Treaty<sup>102</sup> to improve Antarctic and Southern Ocean environmental management, a sea change swept across the Antarctic Treaty System. The CCAMLR embraced the reality that climate change and harvesting were simultaneously threatening the Antarctic ecosystem. In consequence, barriers to establishment of Marine Protected Areas<sup>60</sup> were dismantled. Systematic conservation planning<sup>103</sup>, based

on recognition of evolutionary potential and genetic connectivity, enabled a flexible approach to maintenance of populations and helped achieve the CCAMLR goal of conservation of the Southern Ocean ecosystem.

**Marine ecosystems.** Local change in the marine system continued along trajectories recorded early in the twenty-first century<sup>72–75,78</sup>, with reversals of earlier trends recently becoming apparent. Widely forecast tipping points were averted. New monitoring approaches<sup>104,105</sup> provided opportunities to identify locations where such threshold shifts might be important and to design strategies to avert them, such as temporary ecosystem manipulation<sup>106</sup>. Although ocean acidification continued, the impacts stabilized following the decrease in atmospheric CO<sub>2</sub> levels after 2040. Some population declines were recorded in sensitive species<sup>77</sup>, but others adapted, resulting in less change than was initially forecast<sup>71</sup>. Seal and seabird populations continued to show changes in their foraging range, and changes in body mass and breeding success<sup>62,64</sup>. However, stabilization of sea ice conditions, and a reduction in the frequency of the extreme events to which these species are sensitive<sup>107</sup>, reduced the rate of change. For some species, such as wandering albatross *Diomedea exulans*, weakening in the strength of the westerlies reversed a trend of enhanced breeding success and larger body mass associated with the strengthening westerlies earlier in the century<sup>108</sup>.

**Terrestrial ecosystems.** On land, spread of the two indigenous vascular plant species on the Antarctic Peninsula continued, but at a declining rate by the late 2060s. Changes predicted for the Dry Valleys associated with pulse events<sup>109</sup> continued, but with declining importance after 2060. The Long Term Ecological Research sites proved exceptionally important as a source of evidence to document climate-associated biodiversity responses. Rapid progress in remote sensing techniques<sup>110</sup>, alongside adoption of a suite of essential biodiversity variables, enabled researchers to verify that changes in terrestrial ecosystems were within the bounds anticipated for the low-emissions scenario considered by the IPCC Fifth Assessment Report<sup>1</sup>.

**Invasive species.** The introduction of invasive species along the Antarctic Peninsula initially continued as had been expected. However, none of the world's most invasive species have established, largely because the climate remained inhospitable<sup>83</sup>. Several European and sub-Antarctic species were able to gain a foothold and spread initially, as had been predicted from increasing human activity in the region<sup>82</sup>. Nonetheless, two developments reduced the magnitude of the problem. First, the CEP adopted a systematic, DNA barcoding and web-based surveillance system, which enabled rapid identification of 'unusual' species found by environmental managers, and appropriate action. Second, the parties to the Antarctic Treaty agreed two regulatory frameworks in quick succession in the years leading up to 2030. First, an agreement on the importance of Antarctic genetic resources and bioprospecting was put in place as a further Annex to the Environmental Protocol, thus resolving a decades-long impasse<sup>7</sup>. The renewed focus on the value of indigenous resources improved conservation actions to retain them, including the declaration of a suite of new terrestrial protected areas. Second, a Protocol on Tourism Regulation, initially proposed in the earlier years of the century<sup>5</sup> was eventually agreed as a further Annex to the Environmental Protocol. The Annex regulated all Antarctic activity, recognizing the similarity in impacts of science, tourism and resource extraction activities (the agreement excluded fishing). These developments also prompted further attention to components of diversity not frequently considered, yet of tremendous consequence in the Antarctic terrestrial system—the microbiota<sup>111</sup>. The parties to the Antarctic Treaty agreed by resolution to apply strict biosecurity measures on travel among the Antarctic Conservation Biogeographic Regions<sup>86</sup>. Thus, evolutionary biology continued to document the unusual nature of Antarctic systems and their evolutionary history. The outcomes of this work provided valuable insights into the microbial history of the

planet and the way systems might evolve elsewhere. Moreover, bioprospecting activities delivered new products that improved the health and wellbeing of human populations.

### Change in human engagement with Antarctica

**Effective governance.** Strong action by the international community to mitigate greenhouse gas emissions was echoed in Antarctica, where the parties to the Antarctic Treaty reversed decades of regulatory inaction<sup>5</sup>. Decisive steps were taken to limit the impact of increased human engagement with Antarctica. Motivated by a clearer appreciation of the threats to the region and the global value of better understanding Antarctica and its links to lower latitudes, the parties reaffirmed the commitment to maintain Antarctica as a natural reserve for peace and science<sup>102</sup> and, importantly, strengthened a variety of governance measures to ensure the commitment was put into practice. Although national implementation of measures remained variable, reflecting the complexity of interactions and feedbacks between international law and domestic legislation and politics<sup>5</sup>, the tenor of the international governance conversation had changed.

**International collaboration.** Perhaps most important was the improved relationship between the Antarctic Treaty System and the United Nations, especially through its environment programme. The very cool relations sparked in the 1980s by the 'Question of Antarctica', which remained on the United Nations' General Assembly Agenda until 2005<sup>112</sup>, started to thaw some two decades later. Chief among the catalysts was research showing that Antarctica and the Earth system are inseparable. Moreover, an increasing focus on truly global action to achieve the Sustainable Development Goals<sup>95</sup> helped improve collaboration. Warming relationships between international institutions were also reflected by better collaboration between CCAMLR and other regional fisheries management organizations<sup>5</sup>, to the benefit of the latter. In consequence, regulatory activities across both marine and terrestrial environments improved and an integrated biodiversity strategy<sup>7</sup> facilitated holistic management of the region. These achievements by the Antarctic Treaty System have provided a compelling illustration of the power of effective management of shared international spaces. The governance of Antarctica is now taught in schools worldwide as an example of successful international cooperation in conservation and sustainable resource management.

**Societal benefits.** Human engagement in Antarctica continued to deliver societal benefits on a global scale. New compounds isolated from Antarctic biota are now used in a wide variety of industrial and medical applications (an advance made possible by the Environmental Protocol's new Annex to ensure fair and effective management of intellectual property associated with biological prospecting). The skills of weather, climate and sea level forecasts for the entire Southern Hemisphere now depend heavily on the automatic observing systems established in Antarctica and the Southern Ocean. Lessons learned in redesigning Antarctic bases and logistics to increase efficiency and minimize environmental impact helped accelerate progress to a low-carbon global economy. These advances included improvements in building efficiency, renewable energy and storage, electric vehicles, waste management and autonomous systems.

### Conclusion

We have described two retrospective narratives for Antarctica set in the year 2070: one in which greenhouse gas emissions continued to increase rapidly and little policy action was taken to respond to environmental and social factors affecting Antarctica, and a second in which strong action was taken to reduce emissions and to put in place effective policies to enhance the resilience of Antarctica. The two scenarios are of course highly speculative and intended as counterfactual catalysts for discussion, rather than as predictions of the future. One thing is certain, however: the narrative that eventually plays out will depend substantially on choices made over the next decade<sup>9</sup>.

Antarctica and the Southern Ocean are closely coupled to the rest of the globe. The Antarctic Ice Sheet is the largest and most uncertain potential contributor to future sea level rise. In addition, changes in high southern latitudes will also directly affect the energy budget of the Earth by altering the planetary albedo, the strength of the global overturning circulation, the amount of carbon dioxide in the atmosphere, and the availability of nutrients to support marine life.

Under the high-emissions scenario, Antarctica and the Southern Ocean undergo widespread and rapid change, with global consequences. But the environmental change realized by 2070 will be only a fraction of the change to which we are committed by choices made today, and the rate of change will have increased and continue to accelerate. For example, once initiated, the marine ice sheet instability will result in irreversible loss of large parts of the ice sheet resting on bedrock below sea level. Under the low-emissions scenario, in which global average temperatures remain within 2°C of 1850 values, there is some chance that the buttressing ice shelves will survive and the Antarctic contribution to sea level rise will remain below 1 m. Under the high-emissions scenario, the ice shelves are lost and Antarctica contributes 0.6 m to 3 m of sea level rise by 2300, with an irreversible commitment<sup>45</sup> of 5 m to 9 m, or as much as<sup>33</sup> 15 m in the coming millennia.

Despite the challenges, actions can be taken now that will slow the rate of environmental change, increase the resilience of Antarctica, and reduce the risk of out-of-control consequences. An effective response to the challenges of a changing Antarctica can serve as an example of the power of peaceful international collaboration, as well as demonstrate how integration of physical, biological and social sciences can enable decision-making that is informed by the past and takes account of the long-term consequences of today's choices.

Received: 19 December 2017; Accepted: 13 March 2018;

Published online 13 June 2018.

- Intergovernmental Panel on Climate Change (IPCC) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).  
**The latest comprehensive assessment of the state and future of the climate system, based on observations and Earth system models**
- Frölicher, T. L. et al. Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *J. Clim.* **28**, 862–886 (2015).
- Armour, K. C., Marshall, J., Scott, J. R., Donohoe, A. & Newsom, E. R. Southern Ocean warming delayed by circumpolar upwelling and equatorward transport. *Nat. Geosci.* **9**, 549–554 (2016).
- Sarmiento, J. L., Gruber, N., Brzezinski, M. & Dunne, J. P. High-latitude controls of thermocline nutrients and low latitude biological productivity. *Nature* **427**, 56–60 (2004).
- Dodds, K., Hemmings, A. D. & Roberts, P. *Handbook on the Politics of Antarctica* (Edward Elgar, Cheltenham, 2017).  
**A comprehensive overview of the political arrangements for Antarctica and the Southern Ocean, their current operation, and future challenges**
- Chown, S. L. et al. Challenges to the future conservation of the Antarctic. *Science* **337**, 158–159 (2012).  
**This horizon-scanning assessment provides an inclusive examination of current and future conservation challenges for Antarctica and the Southern Ocean.**
- Chown, S. L. et al. Antarctica and the strategic plan for biodiversity. *PLoS Biol.* **15**, e2001656 (2017).
- Mauritsen, T. & Pincus, R. Committed warming inferred from observations. *Nat. Clim. Change* **7**, 652–655 (2017).
- Rockström, J. et al. A roadmap for rapid decarbonisation. *Science* **355**, 1269–1271 (2017).  
**This paper presents a roadmap to decarbonisation consistent with achieving the goals of the Paris Agreement, highlighting the necessity for fossil fuel emissions to peak by 2020.**
- Drijfhout, S. et al. Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models. *Proc. Natl Acad. Sci. USA* **112**, E5777–E5786 (2015).
- Scheffer, M., Carpenter, S. R., Dakos, V. & van Nes, E. Generic indicators of ecological resilience: inferring the chance of a critical transition. *Annu. Rev. Ecol. Evol. Syst.* **46**, 145–167 (2015).
- Pol, K. et al. Climate variability features of the last interglacial in the East Antarctic EPICA Dome C ice core. *Geophys. Res. Lett.* **41**, 4004–4012 (2014).
- Thompson, D. W. J. et al. Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change. *Nat. Geosci.* **4**, 741–749 (2011).



14. Swart, N. C. & Fyfe, J. C. Observed and simulated changes in the Southern Hemisphere surface westerly wind-stress. *Geophys. Res. Lett.* **39**, L16711 (2012).
15. Durack, P. J., Wijffels, S. E. & Matear, R. J. Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* **336**, 455–458 (2012).
16. Haumann, F. A., Gruber, N., Münnich, M., Frenger, I. & Kern, S. Sea-ice transport driving Southern Ocean salinity and its recent trends. *Nature* **537**, 89–92 (2016).
17. Jacobs, S., Giulivi, C. & Mele, P. Freshening of the Ross Sea during the late 20th century. *Science* **297**, 386–389 (2002).
18. Purkey, S. G. & Johnson, G. C. Warming of global abyssal and deep Southern Ocean waters between the 1990s and 2000s: contributions to global heat and sea level rise budgets. *J. Clim.* **23**, 6336–6351 (2010).
19. van Wijk, E. M. & Rintoul, S. R. Freshening drives contraction of Antarctic Bottom Water in the Australian Antarctic Basin. *Geophys. Res. Lett.* **41**, 1657–1664 (2014).
20. Cai, W. Antarctic ozone depletion causes an intensification of the Southern Ocean super-gyre circulation. *Geophys. Res. Lett.* **33**, L03712 (2006).
21. Sallée, J. B. et al. Assessment of Southern Ocean water mass circulation in CMIP5 models: historical bias and forcing response. *J. Geophys. Res.* **118**, 1830–1844 (2013).
22. Hauri, C., Friedrich, T. & Timmermann, A. Abrupt onset and prolongation of aragonite undersaturation events in the Southern Ocean. *Nat. Clim. Change* **6**, 172–176 (2016).
23. Schmidt, S. et al. Multi-decadal warming of Antarctic waters. *Science* **346**, 1227–1231 (2014).
- This paper summarises changes observed in the Southern Ocean in recent decades, showing that continental shelf waters have warmed in the Amundsen Sea and driven thinning of ice shelves and retreat of grounding lines in this sector.**
24. Mercer, J. H. West Antarctic ice sheet and CO<sub>2</sub> greenhouse effect: a threat of disaster. *Nature* **271**, 321–325 (1978).
25. Favier, L. et al. Retreat of Pine Island Glacier controlled by marine ice-sheet instability. *Nat. Clim. Change* **4**, 117–121 (2014).
26. Joughin, I., Smith, B. E. & Medley, B. Marine ice sheet collapse potentially under way for the Thwaites Glacier basin, West Antarctica. *Science* **344**, 735–738 (2014).
- Ice sheet model simulations suggest that unstable retreat of the Thwaites Glacier, the largest drainage of the West Antarctic Ice Sheet, is already underway, although the timing of full collapse is uncertain.**
27. Rignot, E., Mouginot, J., Morlighem, M., Seroussi, H. & Scheuchl, B. Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011. *Geophys. Res. Lett.* **41**, 3502–3509 (2014).
- Satellite observations show that the grounding lines of the primary glaciers draining the West Antarctic Ice Sheet have retreated over the past two decades.**
28. Phillips, H. A. Surface meltstreams on the Amery Ice Shelf, East Antarctica. *Ann. Glaciol.* **27**, 177–181 (1998).
29. Trusel, L. D. et al. Divergent trajectories of Antarctic surface melt under two twenty-first-century climate scenarios. *Nat. Geosci.* **8**, 927–932 (2015).
30. Kingslake, J., Ely, J. C., Das, I. & Bell, R. E. Widespread movement of meltwater onto and across Antarctic ice shelves. *Nature* **544**, 349–352 (2017).
31. Lenaerts, J. T. M. et al. Meltwater produced by wind-albedo interaction stored in an East Antarctic ice shelf. *Nat. Clim. Change* **7**, 58–62 (2017).
32. Scambos, T., Hulbe, C. & Fahnestock, M. in *Antarctic Peninsula Climate Variability: Historical and Paleoenvironmental Perspectives* (eds Domack, E. et al.) Vol. 79, 79–92 (Antarctic Research Series, AGU, Washington, DC, 2003).
33. DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. *Nature* **531**, 591–597 (2016).
- An ice sheet model, calibrated against sea level records from past warm periods, projects rapid loss of mass from the Antarctic Ice Sheet in response to unmitigated greenhouse gas emissions and a multi-centennial commitment to 15 m of sea level rise from Antarctica.**
34. Paolo, F. S., Fricker, H. A. & Padman, L. Volume loss from Antarctic ice shelves is accelerating. *Science* **348**, 327–331 (2015).
- A multi-mission record (1994 to 2012) of ice-shelf surface height from satellite radar altimetry showed accelerated loss of volume of Antarctica's ice shelves, with early small increases in East Antarctica due to accumulation, and substantial losses in West Antarctica, where some ice shelves thinned by up to 18% in the 18 years.**
35. Robel, A. A. Thinning sea ice weakens buttressing force of iceberg mélange and promotes calving. *Nature Commun.* **8**, 14596 (2017).
36. Rintoul, S. et al. Ocean heat drives rapid basal melt of the Totten Ice Shelf. *Sci. Adv.* **2**, e1601610 (2016).
37. Hellmer, H. H., Kauker, F., Timmermann, R., Determann, J. & Rae, J. Twenty-first-century warming of a large Antarctic ice-shelf cavity by a redirected coastal current. *Nature* **485**, 225–228 (2012).
38. Ross, N. et al. Steep reverse bed slope at the grounding line of the Weddell Sea sector in West Antarctica. *Nat. Geosci.* **5**, 393–396 (2012).
39. Golledge, N., Levy, R. L., McKay, R. & Naish, T. East Antarctic Ice Sheet vulnerable to Weddell Sea warming. *Geophys. Res. Lett.* **44**, 2343–2351 (2017).
40. Liu, Y. et al. Ocean-driven thinning enhances iceberg calving and retreat of Antarctic ice shelves. *Proc. Natl Acad. Sci. USA* **112**, 3263–3268 (2015).
41. Martín-Español, A., Bamber, J. L. & Zammit-Mangion, A. Constraining the mass balance of East Antarctica. *Geophys. Res. Lett.* **44**, 4168–4175 (2017).
42. Medley, B. et al. Temperature and snowfall in western Queen Maud Land increasing faster than climate model projections. *Geophys. Res. Lett.* **45**, 1472–1480 (2018).
43. Nicolas, J. P. & Bromwich, D. H. Climate of West Antarctica and influence of marine air intrusions. *J. Clim.* **24**, 49–67 (2011).
44. Bassis, J. N. & Walker, C. C. Upper and lower limits on the stability of calving glaciers from the yield strength envelope of ice. *Proc. R. Soc. A* **468**, 913–931 (2012).
45. Golledge, N. R. et al. The multi-millennial Antarctic commitment to future sea-level rise. *Nature* **526**, 421–425 (2015).
46. Winkelmann, R., Levermann, A., Ridgwell, A. & Caldeira, K. Combustion of available fossil fuel resources sufficient to eliminate the Antarctic Ice Sheet. *Sci. Adv.* **1**, e1500589 (2015).
47. Feldmann, J. & Levermann, A. Collapse of the West Antarctic Ice Sheet after local destabilization of the Amundsen Basin. *Proc. Natl Acad. Sci. USA* **112**, 14191–14196 (2015).
48. Dutton, A. et al. Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science* **349**, aaa4019 (2015).
49. Clark, P. et al. Consequences of twenty-first-century policy for multi-millennial climate and sea-level change. *Nature Clim. Change* **6**, 360–369 (2016).
50. Hallegatte, S., Green, C., Nicholls, R. J. & Jan Corfee-Morlot, J. Future flood losses in major coastal cities. *Nat. Clim. Change* **3**, 802–806 (2013).
51. Holloway, M. D. et al. Antarctic last interglacial isotope peak in response to sea ice retreat not ice-sheet collapse. *Nature Commun.* **7**, 12293 (2016).
52. Collins, M. A., Brickle, P., Brown, J. & Belchier, M. The Patagonian toothfish: biology, ecology and fishery. *Adv. Mar. Biol.* **58**, 227–300 (2010).
53. Xiong, X. et al. DNA barcoding reveals substitution of Sablefish (*Anoplopoma fimbria*) with Patagonian and Antarctic Toothfish (*Dissostichus eleginoides* and *Dissostichus mawsoni*) in online market in China: how mislabelling opens door to IUU fishing. *Food Control* **70**, 380–391 (2016).
54. Watson, R. A. et al. Global marine yield halved as fishing intensity redoubles. *Fish. Fish.* **14**, 493–503 (2013).
55. Nicol, S. & Foster, J. in *Biology and Ecology of Antarctic Krill* (ed. Siegel, V.) 387–421 (Springer, Cham, 2016).
56. Kawaguchi, S. et al. Risk maps for Antarctic krill under projected Southern Ocean acidification. *Nat. Clim. Change* **3**, 843–847 (2013).
57. Atkinson, A., Siegel, V., Pakhomov, E. A. & Rothery, P. Long-term decline in krill stock and increase in salps within the Southern Ocean. *Nature* **432**, 100–103 (2004).
58. Hofman, R. J. Sealing, whaling and krill fishing in the Southern Ocean: past and possible future effects on catch regulations. *Polar Rec.* **53**, 88–99 (2017).
59. Hill, S. et al. Is current management of the Antarctic krill fishery in the Atlantic sector of the Southern Ocean precautionary? *CCAMLR Sci.* **23**, 31–51 (2016).
60. Brooks, C. M. et al. Science-based management in decline in the Southern Ocean. *Science* **354**, 185–187 (2016).
- Here, an evidence-based argument is laid out for why the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) faces a critical window of opportunity to remain a global leader in resource management.**
61. Trivelpiece, W. Z. et al. Variability in krill biomass links harvesting and climate warming to penguin population changes in Antarctica. *Proc. Natl Acad. Sci. USA* **108**, 7625–7628 (2011).
62. Trathan, P. N. et al. Pollution, habitat loss, fishing, and climate change as critical threats to penguins. *Conserv. Biol.* **29**, 31–41 (2015).
63. Ainley, D. G. et al. How overfishing a large piscine mesopredator explains growth in Ross Sea populations of penguin populations: a framework to better understand impacts of a controversial fishery. *Ecol. Modell.* **349**, 69–75 (2017).
64. Xavier, J. C. et al. Seasonal changes in the diet and feeding behaviour of a top predator indicate a flexible response to deteriorating oceanographic conditions. *Mar. Biol.* **160**, 1597–1606 (2013).
65. Ainley, D. G., Ribic, C. A. & Fraser, W. R. Does prey preference affect habitat choice in Antarctic seabirds? *Mar. Ecol. Prog. Ser.* **90**, 207–221 (1992).
66. Barrera-Oro, E. The role of fish in the Antarctic marine food web: differences between inshore and offshore waters in the southern Scotia Arc and west Antarctic Peninsula. *Antarct. Sci.* **14**, 293–309 (2002).
67. Xavier, J. C., Wood, A. G., Rodhouse, P. G. & Croxall, J. P. Interannual variations in cephalopod consumption by albatrosses at South Georgia: implications for future commercial exploitation of cephalopods. *Mar. Freshw. Res.* **58**, 1136–1143 (2007).
68. St. John, M. A. et al. A dark hole in our understanding of marine ecosystems and their services: perspectives from the mesopelagic community. *Front. Mar. Sci.* **3**, 31 (2016).
69. Krüger, L. et al. Projected distributions of Southern Ocean albatrosses, petrels and fisheries as a consequence of climatic change. *Ecography* **41**, 195–208 (2017).

70. Montes-Hugo, M. et al. Recent changes in phytoplankton communities associated with rapid regional climate change along the Western Antarctic Peninsula. *Science* **323**, 1470–1473 (2009).
71. Deppeler, S. L. & Davidson, A. T. Southern Ocean phytoplankton in a changing climate. *Front. Mar. Sci.* **4**, 40 (2017).
72. Schloss, I. R. et al. Response of phytoplankton dynamics to 19-year (1991–2009) climate trends in Potter Cove (Antarctica). *J. Mar. Syst.* **92**, 53–66 (2012).
73. Fuentes, V. et al. Glacial melting: an overlooked threat to Antarctic krill. *Sci. Rep.* **6**, 27234 (2016).
74. Lynch, H. J., Naveen, R., Trathan, P. N. & Fagan, W. F. Spatially integrated assessment reveals widespread changes in penguin populations on the Antarctic Peninsula. *Ecology* **93**, 1367–1377 (2012).
75. Clucas, G. V. et al. A reversal of fortunes: climate change ‘winners’ and ‘losers’ in Antarctic Peninsula penguins. *Sci. Rep.* **4**, 5024 (2014).
76. Cross, E. L., Peck, L. S. & Harper, E. M. Ocean acidification does not impact shell growth or repair of the Antarctic brachiopod *Liothyrella uva* (Broderip, 1833). *J. Exp. Mar. Biol. Ecol.* **462**, 29–35 (2015).
77. Bednaršek, N. et al. Extensive dissolution of live pteropods in the Southern Ocean. *Nat. Geosci.* **5**, 881–885 (2012).
78. Gutt, J. et al. The Southern Ocean ecosystem under multiple climate change stresses—an integrated circumpolar assessment. *Glob. Change Biol.* **21**, 1434–1453 (2015).
- An integrated assessment of how environmental change drivers will act in concert to affect Southern Ocean benthic, pelagic and sea-ice species and ecosystems.**
79. Lee, J. R. et al. Climate change drives expansion of Antarctic ice-free habitat. *Nature* **547**, 49–54 (2017).
80. Cannone, N., Guglielmin, M., Convey, P., Worland, M. R. & Longo, S. E. F. Vascular plant changes in extreme environments: effects of multiple drivers. *Clim. Change* **134**, 651–665 (2016).
81. Molina-Montenegro, M. A. et al. Occurrence of the non-native annual bluegrass on the Antarctic mainland and its negative effects on native plants. *Conserv. Biol.* **26**, 717–723 (2012).
82. Chown, S. L. et al. Continent-wide risk assessment for the establishment of nonindigenous species in Antarctica. *Proc. Natl Acad. Sci. USA* **109**, 4938–4943 (2012b).
83. Duffy, G. A. et al. Barriers to globally significant invaders are weakening across the Antarctic. *Divers. Distrib.* **23**, 982–996 (2017).
84. Terauds, A. et al. Conservation biogeography of the Antarctic. *Divers. Distrib.* **18**, 726–741 (2012).
85. Hughes, K. A., Pertierra, L. R., Molina-Montenegro, M. A. & Convey, P. Biological invasions in terrestrial Antarctica: what is the current status and can we respond? *Biodivers. Conserv.* **24**, 1031–1055 (2015).
86. Hughes, K. A. & Pertierra, L. Evaluation of non-native species policy development and implementation within the Antarctic Treaty area. *Biol. Conserv.* **200**, 149–159 (2016).
87. Tin, T., Liggett, D., Maher, P. D. & Lamers, M. (eds) *Antarctic Futures. Human Engagement with the Antarctic Environment* (Springer, Dordrecht, 2014).
88. Chown, S. L. & Duffy, G. A. The veiled ecological danger of rising sea levels. *Nat. Ecol. Evol.* **1**, 1219–1221 (2017).
89. Gerland, P. et al. World population stabilization unlikely this century. *Science* **346**, 234–237 (2014).
90. Ng, M. et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **384**, 766–781 (2014).
91. Jacquet, J., Blood-Patterson, E., Brooks, C. & Ainley, D. G. ‘Rational use’ in Antarctic waters. *Mar. Policy* **63**, 28–34 (2016).
92. Haward, M., Jabour, J. & Press, A. J. Antarctic treaty system ready for a challenge. *Science* **338**, 603 (2012).
93. Soga, M. & Gaston, K. J. Extinction of experience: the loss of human–nature interactions. *Front. Ecol. Environ.* **14**, 94–101 (2016).
94. Frieler, K., Mengel, M. & Levermann, A. Delaying future sea-level rise by storing water in Antarctica. *Earth Syst. Dynam.* **7**, 203–210 (2016).
95. United Nations *The Sustainable Development Goals Report 2017*. <https://unstats.un.org/sdgs/files/report/2017/TheSustainableDevelopmentGoalsReport2017.pdf> (UN, 2017).
96. Paolo, F. S. et al. Response of Pacific-sector Antarctic ice shelves to the El Niño/Southern Oscillation. *Nat. Geosci.* **11**, 121–126 (2018).
97. Fogt, R. L., Bromwich, D. H. & Hines, K. M. Understanding the SAM influence on the South Pacific ENSO teleconnection. *Clim. Dyn.* **36**, 1555–1576 (2011).
98. Eyring, V. et al. Sensitivity of 21st century stratospheric ozone to greenhouse gas scenarios. *Geophys. Res. Lett.* **37**, L16807 (2010).
99. Eyring, V. et al. Long-term ozone changes and associated climate impacts in CMIP5 simulation. *J. Geophys. Res. Atmos.* **118**, 5029–5060 (2013).
100. Kuipers Munneke, P., Ligtenberg, S. R. M., van den Broeke, M. R., van Angelen, J. H. & Forster, R. R. Explaining the presence of perennial liquid water bodies in the firn of the Greenland Ice Sheet. *Geophys. Res. Lett.* **41**, 476–483 (2014).
101. Seroussi, H. et al. Continued retreat of Thwaites Glacier, West Antarctica, controlled by bed topography and ocean circulation. *Geophys. Res. Lett.* **44**, 6191–6199 (2017).
102. Antarctic Treaty Consultative Meeting (ATCM) Santiago Declaration on the Twenty Fifth Anniversary of the signing of the Protocol on Environmental Protection to the Antarctic Treaty. [www.ats.aq/documents/ATCM39/ad/atcm39\\_ad003\\_e.pdf](http://www.ats.aq/documents/ATCM39/ad/atcm39_ad003_e.pdf) (ATCM, 2016).
103. Margules, C. R. & Pressey, R. L. Systematic conservation planning. *Nature* **405**, 243–253 (2000).
104. Constable, A. J. et al. Change in Southern Ocean ecosystems I: How changes in physical habitats directly affect marine biota. *Glob. Change Biol.* **20**, 3004–3025 (2014).
105. Rindi, L., Bello, M. D., Dai, L., Gore, J. & Benedetti-Cecchi, L. Direct observation of increasing recovery length before collapse of a marine benthic ecosystem. *Nature Ecol. Evol.* **1**, 0153 (2017).
106. Boyd, P. W. in *Geoengineering Responses to Climate Change: Selected Entries from the Encyclopedia of Sustainability Science and Technology* (eds Lenton, T. & Vaughan, N.) 53–72 (Springer, New York, 2013).
107. Pardo, D., Jenouvrier, S., Weimerskirch, H. & Barbraud, C. Effect of extreme sea surface temperature events on the demography of an age-structured albatross population. *Phil. Trans. R. Soc. Lond. B* **372**, 20160143 (2017).
108. Weimerskirch, H., Louzao, M., de Grissac, S. & Delord, K. Changes in wind pattern alter albatross distribution and life-history traits. *Science* **335**, 211–214 (2012).
109. Andriuzzi, W. S., Adams, B. J., Barrett, J. E., Virginia, R. A. & Wall, D. H. Observed trends of soil fauna in the Antarctic Dry Valleys: early signs of shifts predicted under climate change. *Ecology* **99**, 312–321 (2017).
110. Chown, S. L. et al. The changing form of Antarctic biodiversity. *Nature* **522**, 431–438 (2015).
111. Cavicchioli, R. Microbial ecology of Antarctic aquatic systems. *Nat. Rev. Microbiol.* **13**, 691–706 (2015).
112. Saul, B. & Stephens, S. T. *Antarctica in International Law* (Hart Publishing, Oxford, 2015).
113. Aronson, R. B. et al. No barrier to emergence of bathyal king crabs on the Antarctic shelf. *Proc. Natl Acad. Sci. USA* **112**, 12997–13002 (2015).

**Acknowledgements** This work arose from a panel of Muse Fellows organised as part of a ‘horizon scan’ by the Scientific Committee on Antarctic Research. We acknowledge the contribution of C. Kennicutt, who conceived and led the horizon scan. We also acknowledge the Tinker Foundation for their support of the Tinker–Muse Prize for Science and Policy in Antarctica. J. Matthews and L. Bell drafted the figures. S.R.R. was supported by the Australian Government Cooperative Research Centre (CRC) programme through the Antarctic Climate and Ecosystems CRC, the National Environmental Science Program, and the Centre for Southern Hemisphere Oceans Research (a partnership between CSIRO and the Qingdao National Laboratory for Marine Research). S.L.C. was supported by the Australian Antarctic Science Program. M.H.E. was supported by the Australian Research Council. V.M.D. acknowledges support from Institut Paul Emile Victor and Agence Nationale de la Recherche (ASUMA project number ANR-14-CE01-0001). T.R.N. was supported by a New Zealand Antarctic Research Institute grant and a Royal Society of New Zealand James Cook Fellowship. M.J.S. acknowledges support from the Grantham Foundation for the Protection of the Environment, the UK Natural Environment Research Council and the British Council. J.C.X. was supported by the Foundation for Science and Technology Investigator programme (IF/00616/2013) and the MARE strategic programme (MARE-UID/MAR/04292/2013). R.M.D. was supported by the NSF under award ICER 1664013, and NASA’s Sea Level Rise Program.

**Reviewer information** Nature thanks D. Ainley, K. Dodds and J. Lenaerts for their contribution to the peer review of this work.

**Author contributions** S.R.R. conceived the retrospective narrative approach as a vehicle to highlight the dependence of the future of Antarctica and the Southern Ocean on choices made today. S.R.R. and S.L.C. wrote the initial draft. S.R.R. coordinated the drafting of the paper and developed the concept for the figures. All authors contributed to the discussion of ideas and the writing of the paper.

**Competing interests** S.L.C. is President of the Scientific Committee on Antarctic Research.

**Additional information** Reprints and permissions information is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to S.R.R. **Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Environment and host as large-scale controls of ectomycorrhizal fungi

Sietse van der Linde<sup>1,2,31\*</sup>, Laura M. Suz<sup>2</sup>, C. David L. Orme<sup>1</sup>, Filipa Cox<sup>3</sup>, Henning Andreae<sup>4</sup>, Endla Asi<sup>5</sup>, Bonnie Atkinson<sup>1,2</sup>, Sue Benham<sup>6</sup>, Christopher Carroll<sup>1</sup>, Nathalie Cools<sup>7</sup>, Bruno De Vos<sup>7</sup>, Hans-Peter Dietrich<sup>8</sup>, Johannes Eichhorn<sup>9</sup>, Joachim Gehrmann<sup>10</sup>, Tine Grebenc<sup>11</sup>, Hyun S. Gweon<sup>12,13</sup>, Karin Hansen<sup>14</sup>, Frank Jacob<sup>15</sup>, Ferdinand Kristöfel<sup>16</sup>, Paweł Lech<sup>17</sup>, Miklós Manninger<sup>18</sup>, Jan Martin<sup>19</sup>, Henning Meesenburg<sup>9</sup>, Päivi Merilä<sup>20</sup>, Manuel Nicolas<sup>21</sup>, Pavel Pavlenda<sup>22</sup>, Pasi Rautio<sup>23</sup>, Marcus Schaub<sup>24</sup>, Hans-Werner Schröck<sup>25</sup>, Walter Seidling<sup>26</sup>, Vít Šrámek<sup>27</sup>, Anne Thimonier<sup>24</sup>, Iben Margrete Thomsen<sup>28</sup>, Hugues Titeux<sup>29</sup>, Elena Vanguelova<sup>6</sup>, Arne Verstraeten<sup>7</sup>, Lars Vesterdal<sup>28</sup>, Peter Waldner<sup>24</sup>, Sture Wijk<sup>30</sup>, Yuxin Zhang<sup>1</sup>, Daniel Žlindra<sup>11</sup> & Martin I. Bidartondo<sup>1,2</sup>

**Explaining the large-scale diversity of soil organisms that drive biogeochemical processes—and their responses to environmental change—is critical. However, identifying consistent drivers of belowground diversity and abundance for some soil organisms at large spatial scales remains problematic. Here we investigate a major guild, the ectomycorrhizal fungi, across European forests at a spatial scale and resolution that is—to our knowledge—unprecedented, to explore key biotic and abiotic predictors of ectomycorrhizal diversity and to identify dominant responses and thresholds for change across complex environmental gradients. We show the effect of 38 host, environment, climate and geographical variables on ectomycorrhizal diversity, and define thresholds of community change for key variables. We quantify host specificity and reveal plasticity in functional traits involved in soil foraging across gradients. We conclude that environmental and host factors explain most of the variation in ectomycorrhizal diversity, that the environmental thresholds used as major ecosystem assessment tools need adjustment and that the importance of belowground specificity and plasticity has previously been underappreciated.**

The main projected effects of environmental change on forest processes stem from global and regional perturbations in the carbon (C) and nitrogen (N) cycles<sup>1,2</sup>, and declines in soil biodiversity<sup>3,4</sup>. Globally, mycorrhizal mutualisms mediate soil processes in terrestrial ecosystems<sup>5</sup> and are major drivers of ecosystem carbon and nitrogen dynamics<sup>6</sup>. Soil carbon sequestration<sup>7,8</sup>, tree population dynamics<sup>9</sup> and mitigation of CO<sub>2</sub> fertilization<sup>10</sup> have recently been linked to ectomycorrhizal symbioses, which are ubiquitous drivers of photosynthetic carbon exchange for soil nutrients across temperate and boreal forests<sup>11</sup>.

How changes in ecosystem processes are underpinned by ectomycorrhizal fungi is poorly understood, but probable large-scale effects of those changes—for example, deteriorating tree mineral nutrition and health—are currently being observed<sup>12,13</sup>. Various ecological processes are apparent only at large spatial scales<sup>14</sup>, and there is concern about the lack of baseline ectomycorrhizal distribution data against which to assess the effects of global change<sup>15,16</sup>. Ectomycorrhizal research has emphasized laboratory or local-scale studies that are often reliant on a few species of culturable fungi to provide mechanistic understanding of symbiotic physiology. However, determinants of ectomycorrhizal diversity at local scales are not necessarily their primary drivers at larger scales<sup>17</sup>, and ectomycorrhizal communities are often dominated by

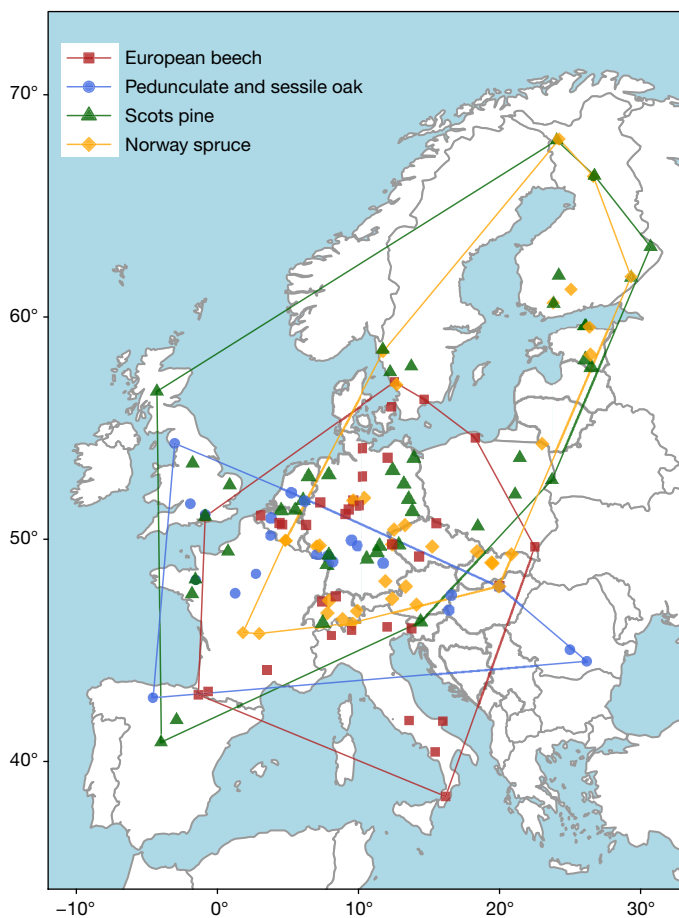
hardly culturable and non-fruiting or inconspicuously fruiting fungi<sup>18</sup>. Furthermore, ectomycorrhizal community composition, richness, fine root biomass and morphology<sup>19–21</sup> and fungal aboveground fruiting<sup>22</sup> indicate different large-scale patterns and responses from plants and animals, and species richness increases with sample area more for ectomycorrhizal fungi than it does for microbes<sup>17,23</sup>. Consequently, there have been repeated calls for unbiased, large-scale, molecular and ecosystem-level baseline data on ectomycorrhizal fungi<sup>15,18,20,24</sup>. Elucidating large-scale ectomycorrhizal diversity is crucial for appropriate experimental design in ecosystem science and model organism selection for experimental and comparative biology<sup>25</sup>.

Unlike multiple local-scale studies in which ectomycorrhizal fungi have been found to be strongly determined by soil environment<sup>26,27</sup>, recent large-scale biogeographical studies report that—other than host identity—soil, climate and atmospheric deposition explain only a limited amount of ectomycorrhizal variability<sup>28–33</sup> (Supplementary Table 1). Even though specialists can be widespread, most ectomycorrhizal fungi are thought to have broad host ranges, although specificity is rarely quantified belowground at large scales<sup>34</sup>.

Current ectomycorrhizal environmental thresholds seldomly integrate occurrence, abundance and directionality of taxon responses,

<sup>1</sup>Life Sciences, Imperial College London, Ascot, UK. <sup>2</sup>Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, London, UK. <sup>3</sup>Earth & Environmental Sciences, University of Manchester, Manchester, UK. <sup>4</sup>Public Enterprise Sachsenforst, Kompetenzzentrum Wald und Forstwirtschaft, Pirna, Germany. <sup>5</sup>Estonian Environment Agency, Tallinn, Estonia. <sup>6</sup>Forest Research, Alice Holt Lodge, Farnham, UK. <sup>7</sup>Nature and Forest Research Institute, Environment and Climate, Geraardsbergen, Belgium. <sup>8</sup>Bavarian State Forestry Institute, Freising, Germany. <sup>9</sup>Northwest German Forest Research Institute, Göttingen, Germany. <sup>10</sup>Landesamt für Natur Umwelt und Verbraucherschutz Nordrhein-Westfalen, Recklinghausen, Germany. <sup>11</sup>Slovenian Forestry Institute, Ljubljana, Slovenia. <sup>12</sup>Biological Sciences, University of Reading, Reading, UK. <sup>13</sup>Centre for Ecology & Hydrology, Wallingford, UK. <sup>14</sup>IVL Swedish Environmental Research Institute, Stockholm, Sweden. <sup>15</sup>Staatsbetrieb Sachsenforst, Pirna, Germany. <sup>16</sup>Federal Research and Training Centre for Forests, Natural Hazards and Landscape (BFW), Wien, Austria. <sup>17</sup>Forest Research Institute, Śękocin Stary, Poland. <sup>18</sup>NARIC Forest Research Institute, Sárovar, Hungary. <sup>19</sup>Landesforstanstalt M-V BT: FVI, Schwerin, Germany. <sup>20</sup>Natural Resources Institute Finland, Oulu, Finland. <sup>21</sup>Office National des Forêts, Recherche-Développement-Innovation, Fontainebleau, France. <sup>22</sup>National Forest Centre, Zvolen, Slovakia. <sup>23</sup>Natural Resources Institute Finland, Rovaniemi, Finland. <sup>24</sup>WSL Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland. <sup>25</sup>Forschungsanstalt für Waldökologie und Forstwirtschaft, Trippstadt, Germany. <sup>26</sup>Thünen Institute of Forest Ecosystems, Eberswalde, Germany. <sup>27</sup>Forestry and Game Management Research Institute, Jiloviště, Czech Republic. <sup>28</sup>Department of Geosciences and Natural Resource Management, University of Copenhagen, Frederiksberg, Denmark. <sup>29</sup>University of Louvain, Earth and Life Institute, Louvain-la-Neuve, Belgium. <sup>30</sup>Swedish Forest Agency, Jönköping, Sweden. <sup>31</sup>Present address: Forest Research, Alice Holt Lodge, Farnham, UK. \*e-mail: [sietse.vanderlinde@forestry.gsi.gov.uk](mailto:sietse.vanderlinde@forestry.gsi.gov.uk)





**Fig. 1** | Map of Europe showing sampled level II plots from the United Nations Economic Commission for Europe International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (UNECE ICP Forests). Polygons depict outer boundaries of the sampled area for each host tree species.

statistical analysis of large-scale standardized datasets or studies of low pollution sites<sup>16,35,36</sup>. Critical loads are essential tools for international atmospheric emissions control<sup>37,38</sup>, but for ectomycorrhizal fungi they differ markedly between Europe and North America<sup>36</sup>. In addition, ectomycorrhizal physiological and morphological plasticity are thought to enhance soil nutrient uptake of trees across environmental gradients<sup>39</sup>; however, foraging-related functional traits are assumed to be fixed at species or genus levels. Wide gradients with abundant observations are needed to link plasticity and environment.

We conducted a detailed mycorrhizal analysis using one of the world's largest and most intensive long-term monitoring networks of soil, atmospheric and vegetation parameters. We analysed 38 variables at 137 plots in 20 European countries across strong environmental gradients. We expected to (1) disentangle significant variability explained by co-varying climatic, soil and atmospheric deposition factors, (2) test the generality of host specificity, (3) detect precise thresholds of mycorrhizal change to inform environmental policy and (4) infer trait plasticity linked to key environmental gradients.

## Results

We examined 29,664 ectomycorrhizas from 9,888 soil cores from 103 plots of approximately 0.25 ha in 18 European countries. We also included data from 34 plots from two previous studies<sup>16,18</sup>, which resulted in a dataset of 39,621 ectomycorrhizas from 137 plots in 20 countries covering an area of approximately 5.5 million km<sup>2</sup> (Fig. 1). After removing short low-quality (12,038), chimaeric (231), non-mycorrhizal (848) and unknown (1,308) internal transcribed spacer DNA sequences, we retained 25,196 sequences that produced 1,406

ectomycorrhizal fungal operational taxonomic units (OTUs), of which 82% were Basidiomycota and 18% were Ascomycota (Fig. 2); 914 OTUs were recorded more than once, and 90% of them were identified to genus or a higher taxonomic level. Of this 90%, 47% were identified to a species level.

## Composition and specificity

We explained 38% of the variance in community composition with forward-selected variables according to the Akaike information criterion. Variables were divided into four partitions: host variables, soil + deposition, climate and geographical distance (Supplementary Table 2). Nine host variables together explained most of the overall community variance (23%), followed by soil + deposition (21%), geographical distance (14%) and climatic variables (12%). The partitions shared 20% of the overall explained variance (Fig. 3).

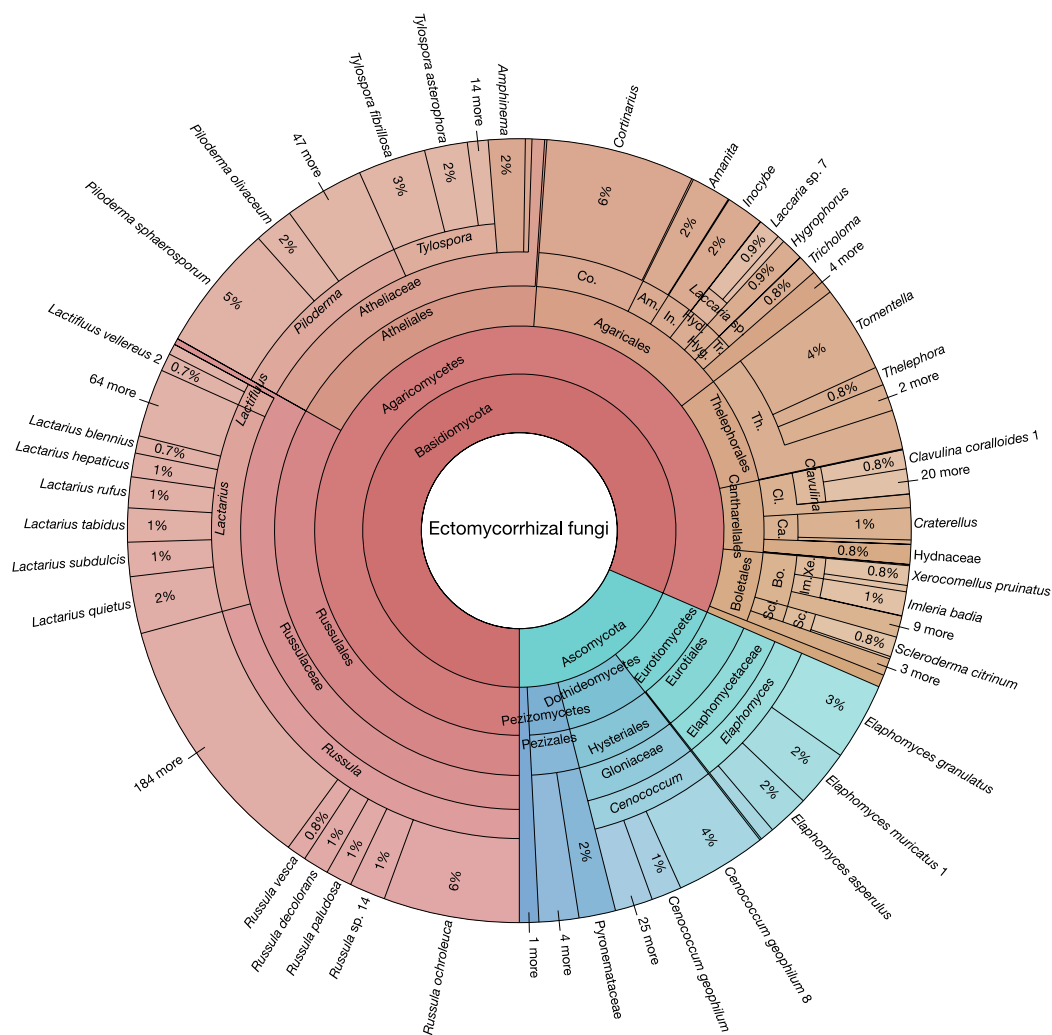
We used global non-metric multidimensional scaling ordinations to visualize ectomycorrhizal fungal community composition, and we fitted environmental variables to the ordination to find the most influential variables (Extended Data Fig. 1 and Extended Data Table 1). Thus, we identified five key variables for subsequent analyses: nitrogen throughfall deposition ( $N_{TFD}$ ), forest-floor pH, mean annual air temperature (MAT), potassium throughfall deposition ( $K_{TFD}$ ) and foliar nitrogen:phosphorus ratio ( $N:P_F$ ).

Almost two-thirds (62%) of ectomycorrhizas correspond to fungi that produce aboveground mushroom-like fruitbodies; the rest produce inconspicuous truffles, crusts or sclerotia. Based on abundance, 48% of ectomycorrhizas were formed by generalists and 52% were specialists to coniferous or broad-leaf hosts. Only 7% of ectomycorrhizas were formed by specialists to one host tree species. Of the 88 OTUs that formed 50 or more ectomycorrhizas, 36 (41%) were generalists and 52 (about 60%) were coniferous or broad-leaf specialists; 11 OTUs (12.5%) were specific to one host species.

## Indicators, thresholds and plasticity

Threshold indicator species analyses identified decreasing ( $z-$ ) and increasing ( $z+$ ) indicator OTUs for all five key environmental variables (Fig. 4 and Extended Data Fig. 2). We identified environmental thresholds of ectomycorrhizal fungal community change by cumulating  $z-$  and  $z+$  change points. For  $N_{TFD}$ , we found a sum( $z-$ ) peak at 5.8 kg nitrogen per ha per year ( $\text{kg N ha}^{-1} \text{ yr}^{-1}$ ) and a sum( $z+$ ) peak at 15.5  $\text{kg N ha}^{-1} \text{ yr}^{-1}$ . For  $N:P_F$ , we detected peaks at 10.2 and 13.3 for sum( $z-$ ) and sum( $z+$ ), respectively. We found a sum( $z-$ ) peak at 6.9  $\text{kg K ha}^{-1} \text{ yr}^{-1}$  and an indistinct sum( $z+$ ) peak at 21.7  $\text{kg K ha}^{-1} \text{ yr}^{-1}$  for  $K_{TFD}$ . There was a distinct peak for forest-floor pH for both sum( $z-$ ) and sum( $z+$ ) at 3.8. Indicator OTUs showed a clear threshold of change for MAT, with a 7.4 °C  $z-$  peak and a distinct 9.1 °C  $z+$  peak. Most  $z-$  OTUs for  $N_{TFD}$ ,  $N:P_F$ , potassium deposition, forest-floor pH and MAT were conifer specialists, whereas all  $z+$  OTUs were generalists or broad-leaf associates. Generally, threshold values based on accumulated change points of individual taxa were less pronounced at the genus than at the OTU level (Extended Data Fig. 3).

The observed frequencies of ectomycorrhizas with emanating hyphae and those with rhizomorphs differed significantly between tree species ( $P < 0.0001$ , d.f. = 3) and soil types ( $P < 0.0001$ , d.f. = 5; Extended Data Table 2a, b); frequencies of ectomycorrhizas with emanating hyphae were higher than expected with beech and spruce, and in Fe–Al soils. Thirty of the 88 most-abundant OTUs ( $\geq 50$  ectomycorrhizas) showed morphological plasticity and 26 of them were also indicators for a key environmental variable. The change in morphology of 17 of these ectomycorrhizal taxa was significantly related to at least one environmental variable (Extended Data Tables 3a, 4a). Morphological plasticity related to at least one variable was found within 12 OTUs when a more stringent 99% sequence similarity was used (Extended Data Tables 3b, 4b). Intraspecific plasticity of individual indicator ectomycorrhizal fungi does not necessarily follow overall community morphological changes, for which logistic regressions showed that mean  $N_{TFD}$  was positively related with hyphal presence ( $P < 0.0001$ ).



**Fig. 2 | Krona chart of taxonomic affiliation of ectomycorrhizal fungi and their relative abundance.** Inner circles represent higher taxonomic ranks, and more detailed taxonomic ranks (up to species level) are presented in outer circles. A full interactive version of this chart is available in Supplementary Fig. 1. Am., Amanitaceae; Bo., Boletaceae; Ca., Cantharellaceae; Cl., Clavulinaceae; Co., Cortinariaceae; Hyd., Hydnangiaceae; Hyg., Hygrophoraceae; Im., *Imleria*; In., Inocybaceae; Sc., *Scleroderma*; Scl., Sclerodermataceae; Th., Thelephoraceae; Tr., Tricholomataceae; Xe., *Xerocomellus*.

There was a negative correlation between hyphal presence and forest-floor pH, N:P<sub>F</sub> and K<sub>TFD</sub>, but no correlation with MAT (Extended Data Table 5). Community-wide, we found negative correlations between rhizomorph presence and all tested environmental variables (Extended Data Table 5).

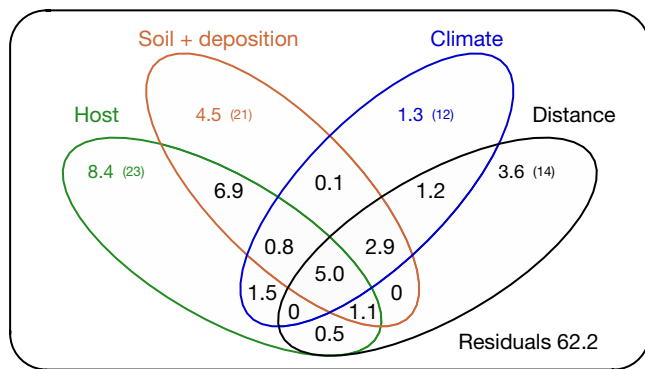
## Discussion

To our knowledge, this is the first large-scale high-resolution study of diversity and distribution of belowground tree symbionts that covers all major European climatic regions for the most abundant tree species. We explain considerable large-scale mycorrhizal diversity with an exceptional range of high-quality environmental, host-related, climatic and geographical variables. We identify large-scale environmental predictors, show the dominance of host specificity, determine environmental indicators and new thresholds of change, and reveal morphological plasticity along environmental gradients. These findings serve as a baseline to assess future change and resilience.

Host-related, soil and atmospheric deposition variables were the most important predictors of ectomycorrhizal community structure across Europe. Four recent large-scale studies<sup>29,31–33</sup> found these variables to be minor predictors, although in local-scale studies, soil environment shows strong effects<sup>26,27</sup>. We distinguished five key environmental variables: N<sub>TFD</sub>, N:P<sub>F</sub>, forest-floor pH, K<sub>TFD</sub> and MAT. Across previous large-scale studies, there is agreement that host species and soil pH are important, but results concerning other variables disagree (Supplementary Table 1). Inconsistent large-scale drivers of diversity and abundance have previously been reported across different microbes<sup>40</sup>, but host is also fundamental for prokaryotes at macroecological scales<sup>41</sup>. Environmental effects on ectomycorrhizal

fungi in previous studies have probably been confounded by: (1) environmental variables from modelled or extrapolated regional sources; (2) non-standardized sampling and spatial pseudo-replication; (3) indirect assignment of mycorrhizal status and traits using databases (for example, UNITE (<https://unite.ut.ee/>), FunGuild (<http://www.stbates.org/guilds/app.php>) or DEEMY (<http://www.deemy.de/>)); (4) semi-quantitative analysis of short DNA sequences; and (5) pooling DNA samples from root hyphae, soil hyphae and dormant propagules even though ectomycorrhizal spore banks differ strongly from active communities on roots at local and large scales<sup>42</sup>, and ephemeral aboveground reproductive structures and soil hyphae correspond weakly with active communities on roots<sup>43,44</sup>. As a result, up to 90% of variation in ectomycorrhizal diversity at large scales has remained unexplained by environmental models<sup>33</sup>. The approach used here is considered more robust<sup>45</sup> and generates higher quality data<sup>46</sup>, but had yet to be scaled up owing to technical challenges. The large unexplained part of community structure may be attributed to factors that were not accounted for, such as disturbance, management history, stochasticity, interactions among variables masking individual effects, measurement and analytical errors, exclusion of rare species, seasonality, using taxonomic instead of functional diversity, and/or not covering complete gradients of each variable across whole geographical ranges of hosts and fungi. In our study, conifers have a larger distribution and thus cover larger environmental gradients; this difference in distribution probably explains the different number of environmental variables linked to community dissimilarities among hosts.

Host-related variables strongly influence ectomycorrhizal fungal communities, thus symbiosis has a major role in shaping ectomycorrhizal distributions. Studies on the host specificity of ectomycorrhizal



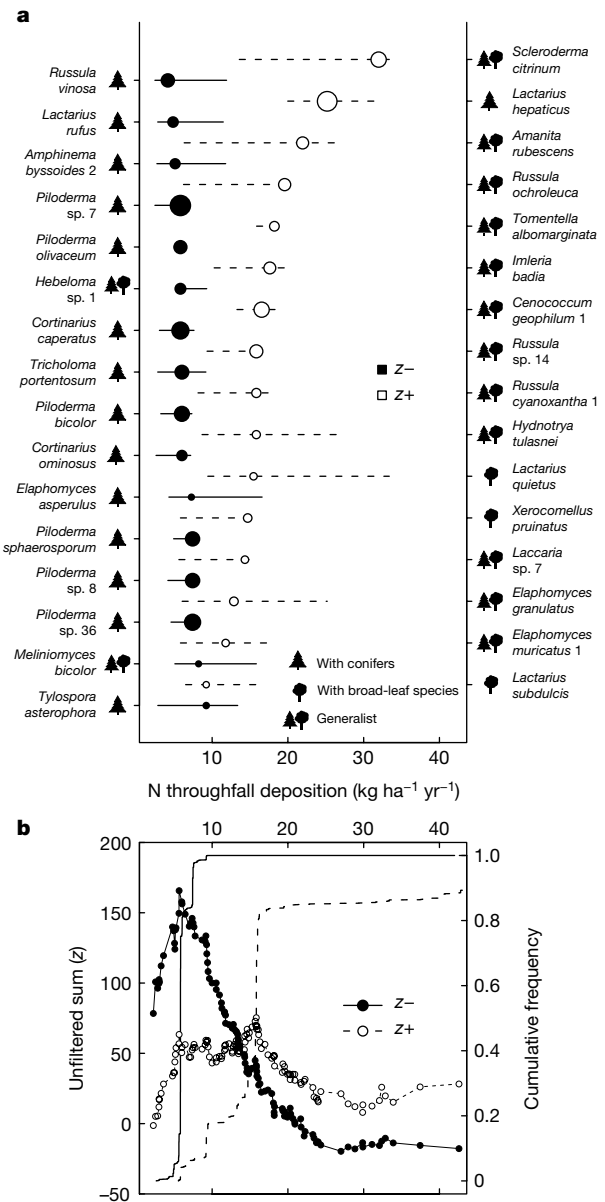
**Fig. 3 | Variation-partitioning Venn diagram.** This diagram shows the percentages of individual contributions of host variables (host species, foliar chemistry and defoliation), soil + deposition variables, climatic variables and geographical distance. The percentage of variance explained by multiple partition models is shown where ellipses overlap. Values in brackets show the total percentage of variance explained by the four partitions. Residual variance represents the percentage unexplained by the four partition models.

fungi at large scales have mainly been based on fruitbody surveys and thus assess specificity on taxonomic rather than abundance levels<sup>47</sup>. Host generalism is considered the rule<sup>48</sup>, but intensive belowground analysis indicates ectomycorrhizal fungal specificity to the most-common European trees matches or exceeds generalism on taxonomic and relative abundance levels, particularly for conifers. We find more conifer specialists and they respond strongly to environmental gradients; the implications of specificity and abundance merit investigation as they can reflect, respectively, more<sup>34,49</sup> and less<sup>50</sup> efficient nutritional mutualisms.

We use threshold indicator taxon analyses for the first time, to our knowledge, for fungi at a continental scale to identify distinct ectomycorrhizal responses to key environmental variables and clear thresholds of change. Indicator species emerged for all key environmental variables, and several ectomycorrhizal taxa were indicators for more than one variable. Different fungi within a family, and even a genus, can be both positive and negative indicators for a variable; for instance, *Thelephora terrestris* and *Tomentella castanea* are negative and positive indicators for N:P<sub>F</sub>, respectively, and *Lactarius rufus* and *Lactarius hepaticus* are negative and positive indicators for N<sub>TFD</sub>, respectively. Nonetheless, genus-level analyses revealed most indicator species patterns hold true at higher taxonomic ranks (Extended Data Fig. 3). In some genera, the aggregate of species acts as indicator, although individual species do not (for example, *Sistotrema*, *Clavulina* and *Boletus* for N<sub>TFD</sub> and K<sub>TFD</sub>). For several genera we find a different response to elevated N<sub>TFD</sub> than previous studies, even those with consistent responses across studies<sup>39</sup> (that is, *Tomentella*, *Tylospora*, *Cenococcum*, *Hebeloma* and *Amanita*). Furthermore, we confirm the response to elevated N<sub>TFD</sub> of several genera that has previously only been recorded in a few studies<sup>39</sup> (that is, *Clavulina*, *Elaphomyces*, *Boletus* and *Amphinema*).

With increasing nitrogen availability, metabolically costly ways of obtaining nitrogen from complex soil organic sources are less cost-effective; fungi that use these pathways (for example, *Cortinarius*, *Piloderma* and *Tricholoma*) are at a disadvantage compared to fungi that use inorganic nitrogen (for example, *Elaphomyces* and *Laccaria*)<sup>39</sup>. Indeed, fungi that use organic nitrogen tended to be negative indicators for nitrogen deposition, and fungi that use inorganic nitrogen tended to be positive indicators.

Some indicator species for K<sub>TFD</sub> are abundant and widespread in Europe (for example, *Elaphomyces asperulus*, *Lactarius quietus* and *Piloderma sphaerosporum*); however, K<sub>TFD</sub> has not been identified as a key variable in previous ectomycorrhizal studies. A meta-analysis showed that in 69% of experiments tree growth responded positively to soil potassium increases<sup>51</sup>, but potassium is highly diffusible in soil and easily accessible for plants. Some K<sub>TFD</sub> may originate



**Fig. 4 | Threshold indicator taxa analyses.** **a**, Individual OTU abundances in response to N<sub>TFD</sub>. Black symbols show taxa declining with increasing N<sub>TFD</sub> (z-), open symbols depict increasing taxa (z+). Symbol size is proportional to magnitude of response (z-score). Horizontal lines represent 5th and 95th quantiles of values resulting in the largest change in taxon z-scores among 1,000 bootstrap replicates. Tree shapes indicate host generalist, conifer- or broad-leaf-specific. **b**, Community-level output of accumulated z-scores per plot is shown in response to N<sub>TFD</sub>.

from canopy leaching; with acidifying pollution, potassium leaches and—if depleted in foliage and litter—potassium availability in soil organic matter could decrease. Moreover, potassium is taken up and translocated by ectomycorrhizal fungi in a specific manner (for example, ectomycorrhizal fungi with hydrophobins transfer less potassium)<sup>11</sup>. This is consistent with our results; most negative indicator genera were hydrophobic and most positive indicator genera were hydrophilic<sup>52</sup>.

Based on the large number of indicator species for MAT, climate should have an important role in shaping ectomycorrhizal communities, as suggested by fruiting phenology studies<sup>53</sup>. However, it is difficult to distinguish MAT from climate and therefore to know whether a fungus occurs somewhere because of prevalent temperatures. Nevertheless, current habitats may become less favourable for many ectomycorrhizal fungi as temperature increases.



Accumulated change-point values of all individual ectomycorrhizal fungi indicate environmental thresholds of change for most key environmental variables. There was a narrow range for fungi negatively affected by  $N_{TFD}$  with a sharp threshold at  $5.8 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ . These mainly conifer specialists thrive in poor soils and pre-industrial nitrogen levels (approximately  $<2 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ ), but cannot keep up with increased  $N_{TFD}$  from industrial, agricultural and transport emissions over the past decades. They are probably outcompeted by fungi that use the additional inorganic nitrogen or avoid additional nitrogen uptake costs<sup>54</sup>, particularly within the temperate distribution ranges of beech and oak where  $N_{TFD}$  is greatest, and organic nitrogen users show some recovery in fruiting if nitrogen pollution decreases<sup>55</sup>. Positively affected fungi—which are mostly host generalists that lack proteolytic abilities—initially do well with additional inorganic N, giving them a competitive advantage. However, the much broader response range and less defined peak at  $15.5 \text{ kg N ha}^{-1} \text{ yr}^{-1}$  for these fungi suggests that adaptation by positively affected fungi to increased  $N_{TFD}$  varies greatly. This might be driven by geographically divergent population-level evolutionary selection pressures on fungi since the Industrial Revolution. Furthermore, naturally enriched microsites (for example, animal latrines, carcasses and disturbances) and macrosites (for example, stands with  $N_2$  fixers) could have pre-adapted certain fungi to high post-industrial levels of nitrogen.

We confirm and extend observations based on fruitbodies and roots at smaller scales<sup>56</sup> that conifer specialists—particularly those with abundant hyphae and rhizomorphs—are more negatively affected by increasing nitrogen than generalists and broad-leaf specialists. The strong differences observed in host specificity between fungi negatively and positively affected by  $N_{TFD}$  may be caused by differences in their enzymatic capability to acquire nitrogen directly from complex soil organic compounds (thus circumventing mineralization) and in resource exchange rate, for example, if specialists transfer more soil nitrogen per unit of tree carbon than generalists<sup>34</sup>. Comparative genetic, physiological and ecological studies of the different sets of dominant indicators are now needed to test alternative models of ectomycorrhizal community optimization versus parasitism under changing carbon and nitrogen conditions<sup>57</sup> through species replacement, plasticity and/or evolution<sup>58</sup>.

Large-scale belowground analysis contributes important information on ecosystem assessment tools for a uniquely important guild of forest organisms. Critical loads for eutrophying nitrogen deposition were previously estimated for ectomycorrhizal fungi at  $5\text{--}10 \text{ kg N ha}^{-1} \text{ yr}^{-1}$  for North America<sup>36</sup> and  $10\text{--}20 \text{ kg N ha}^{-1} \text{ yr}^{-1}$  for Europe<sup>59</sup>, largely based on expert opinion and aboveground data. Thresholds based on European ectomycorrhizal data have focused on a few sites across smaller gradients, or ectomycorrhizal richness and evenness instead of community composition<sup>16,35</sup>. Our large nitrogen-deposition gradient leads to a much lower European threshold value for a substantial ectomycorrhizal shift at  $5\text{--}6 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ , based on both throughfall and open field deposition data, approaching recent lower estimates for other forest organisms<sup>60,61</sup>. Caution is needed when inferring absolute values for critical loads, but based on our results the critical loads for European forests require adjustment towards those for North American forests, and ectomycorrhizal and forest change thresholds need aligning to explain the alarming deterioration in European tree nutrition<sup>13</sup>. Critical  $N:P_F$  are considered to be plant specific<sup>62</sup> and  $N:P_F$  has been linked to tree health, with breakpoint values of 7.3 for conifers and 14.8 for broad-leaf trees with respect to defoliation<sup>12</sup>. We show that lower (10.2) and upper (13.3)  $N:P_F$  thresholds for ectomycorrhizal communities are linked to conifers and broad-leaf trees, respectively. Community threshold forest-floor pH levels for negative and positive indicator species overlap. Although soil pH is anthropogenically influenced (for example, through liming) and soil acidification affects parts of Europe<sup>63</sup>, the major soil pH differences across forests arise from soil parent material and climatic differences over long timescales, and must have long been an influence on ectomycorrhizal communities. Nonetheless, individual species could be affected by changing pH levels.

For  $K_{TFD}$ , no threshold values for ectomycorrhizal composition have yet been published. We identify a  $5\text{--}8 \text{ kg K ha}^{-1} \text{ yr}^{-1}$  threshold for declining species; however,  $K_{TFD}$  results partly from potassium uptake and leaching by trees, which may be influenced by ectomycorrhizal fungi themselves. Therefore, research into potassium deposition and cycling is needed for ectomycorrhizal communities<sup>11</sup> and forests<sup>51</sup>.

Physiological and morphological heterogeneity and plasticity of ectomycorrhizal mycelium have previously been considered to be responsible for enabling trees to rapidly take up soil nutrients<sup>64,65</sup>; here we show morphological plasticity within dominant ectomycorrhizal taxa and changes over environmental gradients. This has important implications for functional diversity studies at large scales and/or across gradients. Indirect assignment of ectomycorrhizal functional traits to taxonomic groups merits caution and temporal variation of these traits merits investigation.

We conclude that intensive and extensive organismal and environmental data collection, with multiple biotic and abiotic co-varying factors, reveals that soil, atmospheric deposition and climate variables control large-scale patterns of species distributions in ectomycorrhizal communities. Such data enable the linking of species and community responses to environmental thresholds acting across macroecological scales and deliver new insights into spatial variation in specificity and functional trait plasticity below the ground.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0189-9>.

Received: 29 October 2017; Accepted: 2 May 2018;

Published online: 06 June 2018

- Canadell, J. G. et al. Contributions to accelerating atmospheric  $\text{CO}_2$  growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proc. Natl Acad. Sci. USA* **104**, 18866–18870 (2007).
- Galloway, J. N. et al. Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science* **320**, 889–892 (2008).
- Commission of the European Communities. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions – Thematic Strategy for Soil Protection (COM(2006) 231) <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52006DC0231> (2006).
- Janssens, I. A. et al. Reduction of forest soil respiration in response to nitrogen deposition. *Nat. Geosci.* **3**, 315–322 (2010).
- Johnson, N. C. & Jansa, J. in *Mycorrhizal Mediation of Soil: Fertility, Structure, and Carbon Storage* (eds Johnson, N. C. et al.) 1–6 (Elsevier, Amsterdam, 2017).
- van der Heijden, M. G. A., Martin, F. M., Selosse, M.-A. & Sanders, I. R. Mycorrhizal ecology and evolution: the past, the present, and the future. *New Phytol.* **205**, 1406–1423 (2015).
- Averill, C., Turner, B. L. & Finzi, A. C. Mycorrhiza-mediated competition between plants and decomposers drives soil carbon storage. *Nature* **505**, 543–545 (2014).
- Clemmensen, K. E. et al. Roots and associated fungi drive long-term carbon sequestration in boreal forest. *Science* **339**, 1615–1618 (2013).
- Bennett, J. A. et al. Plant-soil feedbacks and mycorrhizal type influence temperate forest population dynamics. *Science* **355**, 181–184 (2017).
- Terrer, C., Vicca, S., Hungate, B. A., Phillips, R. P. & Prentice, I. C. Mycorrhizal association as a primary control of the  $\text{CO}_2$  fertilization effect. *Science* **353**, 72–74 (2016).
- Smith, S. E. & Read, D. E. *Mycorrhizal Symbiosis* 3rd edn (Academic, London, 2008).
- Veresoglou, S. D. et al. Exploring continental-scale stand health – N:P ratio relationships for European forests. *New Phytol.* **202**, 422–430 (2014).
- Jonard, M. et al. Tree mineral nutrition is deteriorating in Europe. *Glob. Change Biol.* **21**, 418–430 (2015).
- Levin, S. A. Multiple scales and the maintenance of biodiversity. *Ecosystems* **3**, 498–506 (2000).
- Lilleskov, E. A. & Parrent, J. L. Can we develop general predictive models of mycorrhizal fungal community-environment relationships? *New Phytol.* **174**, 250–256 (2007).
- Suz, L. M. et al. Environmental drivers of ectomycorrhizal communities in Europe's temperate oak forests. *Mol. Ecol.* **23**, 5628–5644 (2014).
- Peay, K. G. & Matheny, P. B. in *Molecular Mycorrhizal Symbiosis* (ed. Martin, F.) 341–361 (John Wiley & Sons, Hoboken, 2016).
- Cox, F., Barsoum, N., Lilleskov, E. A. & Bidartondo, M. I. Nitrogen availability is a primary determinant of conifer mycorrhizas across complex environmental gradients. *Ecol. Lett.* **13**, 1103–1113 (2010).

19. Cudlin, P. et al. Fine roots and ectomycorrhizas as indicators of environmental change. *Plant Biosyst.* **141**, 406–425 (2007).
20. Tedersoo, L. et al. Towards global patterns in the diversity and community structure of ectomycorrhizal fungi. *Mol. Ecol.* **21**, 4160–4170 (2012).
21. Ostonen, I. et al. Adaptive root foraging strategies along a boreal-temperate forest gradient. *New Phytol.* **215**, 977–991 (2017).
22. Kausrud, H. et al. Warming-induced shift in European mushroom fruiting phenology. *Proc. Natl Acad. Sci. USA* **109**, 14488–14493 (2012).
23. Peay, K. G., Bruns, T. D., Kennedy, P. G., Bergemann, S. E. & Garbelotto, M. A strong species–area relationship for eukaryotic soil microbes: island size matters for ectomycorrhizal fungi. *Ecol. Lett.* **10**, 470–480 (2007).
24. Peay, K. G., Bidartondo, M. I. & Arnold, A. E. Not every fungus is everywhere: scaling to the biogeography of fungal–plant interactions across roots, shoots and ecosystems. *New Phytol.* **185**, 878–882 (2010).
25. Suz, L. M. et al. Monitoring ectomycorrhizal fungi at large scales for science, forest management, fungal conservation and environmental policy. *Ann. For. Sci.* **72**, 877–885 (2015).
26. Peay, K. G., Kennedy, P. G., Davies, S. J., Tan, S. & Bruns, T. D. Potential link between plant and fungal distributions in a dipterocarp rainforest: community and phylogenetic structure of tropical ectomycorrhizal fungi across a plant and soil ecotone. *New Phytol.* **185**, 529–542 (2010).
27. Taylor, D. L. et al. A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecol. Monogr.* **84**, 3–20 (2014).
28. Bahram, M., Peay, K. G. & Tedersoo, L. Local-scale biogeography and spatiotemporal variability in communities of mycorrhizal fungi. *New Phytol.* **205**, 1454–1463 (2015).
29. Kennedy, P. G., Garibay-Orijel, R., Higgins, L. M. & Angeles-Arguiz, R. Ectomycorrhizal fungi in Mexican *Alnus* forests support the host co-migration hypothesis and continental-scale patterns in phylogeography. *Mycorrhiza* **21**, 559–568 (2011).
30. Kennedy, P. G. et al. Scaling up: examining the macroecology of ectomycorrhizal fungi. *Mol. Ecol.* **21**, 4151–4154 (2012).
31. Pöhlme, S. et al. Biogeography of ectomycorrhizal fungi associated with alders (*Alnus* spp.) in relation to biotic and abiotic variables at the global scale. *New Phytol.* **198**, 1239–1249 (2013).
32. Talbot, J. M. et al. Endemism and functional convergence across the North American soil mycobiome. *Proc. Natl Acad. Sci. USA* **111**, 6341–6346 (2014).
33. Tedersoo, L. et al. Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).
34. Molina, R. & Horton, T. R. in *Mycorrhizal Networks* (ed. Horton, T. R.) 1–39 (Springer Science + Business Media Dordrecht, Dordrecht, 2015).
35. de Witte, L. C., Rosenstock, N. P., van der Linde, S. & Braun, S. Nitrogen deposition changes ectomycorrhizal communities in Swiss beech forests. *Sci. Total Environ.* **605–606**, 1083–1096 (2017).
36. Pardo, L. H. et al. Effects of nitrogen deposition and empirical nitrogen critical loads for ecoregions of the United States. *Ecol. Appl.* **21**, 3049–3082 (2011).
37. Hettelingh, J.-P. et al. in *Critical Loads and Dynamic Risk Assessments: Nitrogen, Acidity and Metals in Terrestrial and Aquatic Ecosystems* (eds de Vries, W. et al.) 613–635 (Springer Science + Business Media Dordrecht, Dordrecht, 2015).
38. Reis, S. et al. From acid rain to climate change. *Science* **338**, 1153–1154 (2012).
39. Lilleskov, E. A., Hobbie, E. A. & Horton, T. R. Conservation of ectomycorrhizal fungi: exploring the linkages between functional and taxonomic responses to anthropogenic N deposition. *Fung. Ecol.* **4**, 174–183 (2011).
40. Hendershot, J. N., Read, Q. D., Henning, J. A., Sanders, N. J. & Classen, A. T. Consistently inconsistent drivers of microbial diversity and abundance at macroecological scales. *Ecology* **98**, 1757–1763 (2017).
41. Thompson, L. R. et al. A natural catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
42. Glassman, S. I. et al. A continental view of pine-associated ectomycorrhizal fungal spore banks: a quiescent functional guild with a strong biogeographic pattern. *New Phytol.* **205**, 1619–1631 (2015).
43. Gardes, M. & Bruns, T. D. Community structure of ectomycorrhizal fungi in a *Pinus muricata* forest: above- and below-ground views. *Can. J. Bot.* **74**, 1572–1583 (1996).
44. Anderson, I. C. & Cairney, J. W. G. Ectomycorrhizal fungi: exploring the mycelial frontier. *FEMS Microbiol. Rev.* **31**, 388–406 (2007).
45. Buée, M., Sentauba, E. & Murat, C. in *Molecular Mycorrhizal Symbiosis* (ed. Martin, F.) 323–406 (John Wiley & Sons, Hoboken, 2016).
46. Tedersoo, L. & Nilsson, R. H. in *Molecular Mycorrhizal Symbiosis* (ed. Martin, F.) 299–322 (John Wiley & Sons, Hoboken, 2016).
47. Newton, A. C. & Haigh, J. M. Diversity of ectomycorrhizal fungi in Britain: a test of the species–area relationship, and the role of host specificity. *New Phytol.* **138**, 619–627 (1998).
48. Peay, K. G. The mutualistic niche: mycorrhizal symbiosis and community dynamics. *Annu. Rev. Ecol. Syst.* **47**, 143–164 (2016).
49. Taylor, A. F. S., Fransson, P. M., Högborg, P., Högborg, M. N. & Plamboeck, A. H. Species level patterns in  $^{13}\text{C}$  and  $^{15}\text{N}$  abundance of ectomycorrhizal and saprotrophic fungal sporocarps. *New Phytol.* **159**, 757–774 (2003).
50. Hortal, S. et al. Role of plant–fungal nutrient trading and host control in determining the competitive success of ectomycorrhizal fungi. *ISME J.* **11**, 2666–2676 (2017).
51. Tripler, C. E., Kaushal, S. S., Likens, G. E. & Walter, M. T. Patterns in potassium dynamics in forest ecosystems. *Ecol. Lett.* **9**, 451–466 (2006).
52. Agerer, R. Exploration types of ectomycorrhizae. *Mycorrhiza* **11**, 107–114 (2001).
53. Boddy, L. et al. Climate variation effects on fungal fruiting. *Fung. Ecol.* **10**, 20–33 (2014).
54. Wallander, H. A new hypothesis to explain allocation of dry matter between mycorrhizal fungi and pine seedlings in relation to nutrient supply. *Plant Soil* **168**, 243–248 (1995).
55. van Strien, A. J., Boomsliuter, M., Noordeloos, M. E., Verweij, R. J. T. & Kuyper, T. W. Woodland ectomycorrhizal fungi benefit from large-scale reduction in nitrogen deposition in the Netherlands. *J. Appl. Ecol.* **55**, 290–298 (2018).
56. Arnolds, E. Decline of ectomycorrhizal fungi in Europe. *Agric. Ecosyst. Environ.* **35**, 209–244 (1991).
57. Lilleskov, E. A. in *The Fungal Community* (eds Dighton, J. et al.) 769–801 (CRC, Boca Raton, 2005).
58. Kiers, T. E., Palmer, T. M., Ives, A. R., Bruno, J. F. & Bronstein, J. L. Mutualisms in a changing world: an evolutionary perspective. *Ecol. Lett.* **13**, 1459–1474 (2010).
59. Bobbink, R. & Hettelingh, J.-P. in *Review and Revision of Empirical Critical Loads and Dose-Response Relationships (RIVM Report 680359002)* (eds Bobbink, R. & Hettelingh, J.-P.) 135–171 (Coordination Centre for Effects, National Institute for Public Health and the Environment (RIVM), Bilthoven, 2011).
60. Giordani, P. et al. Detecting the nitrogen critical loads on European forests by means of epiphytic lichens. A signal-to-noise evaluation. *For. Ecol. Manage.* **311**, 29–40 (2014).
61. Leppänen, S. M., Salemaa, M., Smolander, A., Mäkipää, R. & Tirola, M. Nitrogen fixation and methanotrophy in forest mosses along a N deposition gradient. *Environ. Exp. Bot.* **90**, 62–69 (2013).
62. Güsewell, S. N. Pratiens in terrestrial plants: variation and functional significance. *New Phytol.* **164**, 243–266 (2004).
63. Cools, N. & De Vos, B. Availability and evaluation of European forest soil monitoring data in the study on the effects of air pollution on forests. *iForest* **4**, 205–211 (2011).
64. Hazard, C. & Johnson, D. Does genotypic and species diversity of mycorrhizal plants and fungi affect ecosystem function? *New Phytol.* <https://doi.org/10.1111/nph.15010> (2018).
65. Chen, W. et al. Root morphology and mycorrhizal symbioses together shape nutrient foraging strategies of temperate trees. *Proc. Natl Acad. Sci. USA* **113**, 8741–8746 (2016).

**Acknowledgements** We acknowledge funding from NERC grant NE/K006339/1 to M.I.B. and C.D.L.O. Analysis was partly based on the ICP Forests PCC Database (<http://icp-forests.net>). ICP Forests FSCC provided the first level II soil survey data. ICP Forests PCC and observers, technicians and scientists performed long-term sampling, analyses and environmental data handling largely funded by national institutions and ministries, supported by governmental bodies, services and landowners, and partially EU-funded under Regulation (EC) No. 2152/2003 (Forest Focus), project LIFE07ENV/D/000218 (FutMon), and through SWETHRO. Co-financing for D.Ž. and T.G. was provided by P4-0107 (RS Higher Education, Science and Technology Ministry). We thank D. Devey and L. Csiba for laboratory assistance; S. Boersma, F. van der Linde, H. van der Linde, J. van der Linde, C. Gonzales, A. Lenz, R. Lenz, S. Wipf, L. Garfoot, B. Spake, W. Rimington, J. Kowal, T. Solovieva, D. Gane, M. Terrington, J. Alden, A. Otway, V. Kemp, M. Edgar, Y. Lin, A. Drew, E. Booth, P. Cachera, R. De-Kayne, J. Downie, A. Tweedy, E. Moratto, E. Ek, P. Helminen, R. Lievonen, P. Närhi, A. Ryyänänen, M. Rupel, J. Draing and F. Heun for field and laboratory work; R. Castilho for bioinformatics; K.-H. Larsson, P.-A. Moreau, J. Nuytinck and M. Ryberg for taxonomy; and N. Barsoum, E. Lilleskov, D. Read and T. Kuyper for discussions throughout.

**Reviewer information** *Nature* thanks A. Dahlberg, P. Kennedy, F. Teste and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** M.I.B. conceived the study. S.v.d.L., M.I.B., F.C., L.M.S. and B.A. led most sampling design and fieldwork. S.v.d.L., B.A., L.M.S., F.C., Y.Z. and M.I.B. processed and analysed samples. H.A., E.A., S.B., N.C., B.D.V., H.-P.D., J.E., J.G., T.G., K.H., F.J., F.K., P.L., M.M., J.M., H.M., P.M., M.N., P.P., P.R., M.S., H.-W.S., W.S., V.S., A.T., I.M.T., H.T., E.V., A.V., L.V., P.W., S.W. and D.Ž. assisted with fieldwork and collected, collated and validated long-term environmental data. S.v.d.L., H.S.G. and C.D.L.O. performed bioinformatics. S.v.d.L., C.D.L.O. and L.M.S. performed data analysis. C.C. summarized literature. S.v.d.L. drafted the manuscript, M.I.B. provided chief contributions, and C.D.L.O. and L.M.S. contributed extensively. All authors wrote and reviewed the manuscript. S.v.d.L., L.M.S., C.D.L.O. and M.I.B. led revision of the manuscript.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0189-9>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0189-9>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to S.v.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

No statistical methods were used to predetermine sample size.

**Sampling and processing.** Since 1995, the International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (<http://icp-forests.net/>)<sup>66</sup> has been intensively monitoring about 800 plots (level II) in major forest ecosystems across Europe<sup>67</sup>. Their extensive in situ data better reflect the local environmental conditions of plots than regional modelled or extrapolated data<sup>68</sup>. These level II plots of at least 0.25 ha and located within homogenous forest stands are structurally diverse and cover a representative mixture of European managed forest types (ranging from plantations to natural regenerating forests)<sup>69</sup>. European forests are dominated by Scots pine, Norway spruce and European beech (60% of the European Union forest area), with the next three most-common tree species together covering 10%. We selected all ICP Forests level II plots in which deposition, meteorology, foliar chemistry, soil and preferably soil solution data are measured simultaneously, and between September 2013 and September 2015 we sampled plots with European beech (*Fagus sylvatica* L.;  $n = 35$ ), Norway spruce (*Picea abies* (L.) H. Karst;  $n = 36$ ) or Scots pine (*Pinus sylvestris* L.;  $n = 32$ ) as the dominant (>50% abundance) tree species. We combined these with previously published additional data that were similarly collected from Scots pine level II plots<sup>18</sup> ( $n = 12$ ) and pedunculate and sessile oak (*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl)<sup>16</sup> ( $n = 22$ ), to give a widespread coverage of European forest areas (Fig. 1).

We used Sanger DNA sequencing of the full internal transcribed spacer (ITS) amplicon from individual ectomycorrhizas to maximize resolution of identifications, obtain relative abundance data and link DNA sequences directly to morphology, following previously published standardized sampling protocols<sup>16,18</sup>. In brief, on each plot ( $n = 137$ ) 24 trees of the investigated target tree species were randomly selected and from those trees a transect was made to the nearest tree of the target species, then four soil samples (25-cm deep, 2-cm diameter) were collected at equal distances on each transect. When plots contained multiple tree species, areas with non-target tree species were avoided. Soil samples were stored at 4 °C for up to ten days until they were processed. Roots from each soil core were rinsed on a 0.5-mm sieve, and mycorrhizal roots were collected for five minutes using a dissecting microscope. Subsequently, from each soil sample, an individual mycorrhiza was sampled from the three longest roots, resulting in 288 mycorrhizas per plot. Morphological characteristics of each mycorrhiza were recorded, including presence/absence of emanating hyphae and rhizomorphs, and turgor to assess activity. Genomic DNA from individual mycorrhizas was obtained using Extract-N-Amp (Sigma-Aldrich), and the ITS region of the nuclear rDNA was amplified using ITS1F<sup>70</sup> and ITS4<sup>71</sup> primers. Amplicons were purified using ExoSAP-IT (USB) and sequenced bidirectionally using BigDye3.1 with an ABI 3730 DNA Analyzer (Applied Biosystems).

**Environmental data.** On the level II plots various environmental long-term measurements (average 14 years) were carried out using national protocols based on a harmonized methodology<sup>72</sup> (see Supplementary Table 2). Soil types were classified in ten types: Andosols, Arenosols, Calcisols, Cambisols, Leptosols, Podzols, Regosols, Umbrisols, soil types characterized by an Argic B horizon (that is, Luvisols and Alisols), and soils with gleyic properties (that is, Gleysols and Stagnosols)<sup>63,73</sup>. While maximizing the number of plots without missing values ( $n = 108$ ), we selected available data—including forest age, level of defoliation<sup>74</sup>, geographical coordinates and elevation—along with soil (eight variables) and foliar (seven variables of investigated tree species)<sup>75</sup> data, atmospheric throughfall deposition chemistry (wet and dry under forest canopy deposition, eleven variables)<sup>76</sup> and meteorology (six variables)<sup>77</sup>.

**Bioinformatics.** We used Phred<sup>78</sup> to obtain base quality scores ( $Q$ ) for both forward and reverse DNA sequences from all individual mycorrhizas, including previously published DNA sequences<sup>16,18</sup>. The two sequences obtained from each mycorrhiza were assembled in Geneious (version 8.1.8)<sup>79</sup>, with the De novo Assemble tool. We used Trimmomatic<sup>80</sup> to remove low-quality bases ( $Q < 20$ ) at either end of the sequences and then discarded short reads (<100 remaining bp). We then used the uclust\_ref tool in vsearch<sup>81</sup> to match chimaeric sequences against the UNITE reference database (version 7.1, 22/08/2016).

We used the usearch\_global tool in vsearch to identify remaining DNA sequences with a percentage match  $\geq 97\%$  to UNITE 7.1 species hypotheses<sup>82</sup>. From the remaining unmatched sequences, we first removed all sequences with ambiguous base pair codes and then used the cluster\_fast tool in vsearch, to identify de novo OTU clusters. The unmatched sequences were then matched to the centroids of these de novo clusters; sequences were accepted with a percentage identity  $\geq 97\%$  and the remaining sequences were discarded.

We used three sources of information for each de novo centroid to confirm the identification of the fungal sequences and to provide tentative classifications. First, we examined the ten best alignments from BLAST searches<sup>83</sup> of the GenBank nucleotide database. Second, we trained RDP Classifier<sup>84</sup> against the UNITE 7.1 database and then classified the de novo centroids against the trained database.

Third, we used vsearch to obtain the best match of each centroid to the UNITE 7.1 species hypotheses.

Finally, we checked the ectomycorrhizal status of all OTUs by comparing the taxonomic classification based on UNITE with the literature<sup>85,86</sup>. When OTUs assigned in UNITE to a species hypothesis were identified to a taxonomic level that includes both ectomycorrhizal and non-ectomycorrhizal fungi (for example, Agaricomycetes species), we retrieved the taxonomic names associated with all UNITE DNA sequences within that species hypothesis to assess the level of uncertainty in the classification of the species hypothesis. We discarded de novo OTUs with less-resolved classification: (1) those with a classification that was distant from known ectomycorrhizal fungi, (2) those where the root tip morphology suggested possibly dead plant or fungal tissue, and (3) those which were based on relatively short sequences (<150 bp). The set of identified ectomycorrhizal fungal sequences was then used to construct an abundance matrix of OTUs across sites. We used the Hellinger transformation of proportion abundance<sup>87</sup> in subsequent analyses. Host specificity of abundant OTUs ( $\geq 50$  ectomycorrhizas) was established by scoring occurrence at plots with the different tree hosts. The OTUs occurring with one host tree species in a plot were considered strictly specific and OTUs occurring with both coniferous and broad-leaf species, or with more than two tree species were considered generalists.

**Statistical analysis.** We used R (version 3.3.3) for statistical analyses and generating figures<sup>88</sup>.

To quantify the importance of host variables, soil and deposition chemistry, climate and geographical distance on ectomycorrhizal fungal community composition, variances were partitioned following previous publications<sup>89,90</sup>. Explanatory variables describing plot and tree characteristics were grouped in the following partitions: (1) host (host species, foliar chemistry and defoliation), (2) soil and deposition chemistry (soil characteristics and throughfall deposition), (3) climate (climatic region, MAT, precipitation, growing season length, minimum and maximum annual temperatures, and elevation) and (4) geographical distance (excluding elevation). The most-relevant variables in each partition were found through forward-selection model-building with the redundancy analysis method based on the Akaike information criterion and  $P < 0.05$  using ordistep in the vegan package<sup>91</sup>. Geographical distances are the great circle distances, calculated using the mean Earth radius between the minimum and maximum latitude of plots in this study ( $r = 6,365$  km) with rdist.earth in the fields package<sup>92</sup>. Great circle distances are commonly used in large-scale macroecological studies to approach real distances between sampling sites<sup>93,94</sup>. The geographical distance matrix was transformed to rectangular data by extracting spatial vectors with principal coordinates of neighbour matrices (PCNM) using pcnm (vegan). To build the geographical distance model, PCNM vectors accounting for autocorrelation were extracted ( $P < 0.05$ ) using MoranI (ltools package)<sup>95</sup> and forward selected. Variation partitioning was carried out for the 108 plots with the selected environmental data using varpart (vegan).

Global non-metric multi-dimensional scaling ordinations were used to explore and visualize the main factors that affect ectomycorrhizal fungal community composition with metaMDS (vegan). Environmental variables (Supplementary Table 2) were fitted to the ordination plots using envfit (vegan). Ordinations were performed for the 108 plots with the selected environmental data. To limit co-linearity effects between variables, we selected key environmental variables from the envfit results with  $R^2 > 0.4$  and  $P < 0.01$ . In case of correlations ( $r \geq 0.7$ ) between those variables, the most-commonly measured environmental variable (Supplementary Table 1) was selected: N<sub>TFD</sub>, forest-floor pH, MAT, K<sub>TFD</sub> and N:P<sub>F</sub>.

Indicator species for the key environmental variables were detected and their threshold values were calculated using threshold indicator species analyses (TITAN2)<sup>96</sup>. The sums of the indicator species scores of all OTUs were used to detect lower and upper ectomycorrhizal community thresholds for key environmental variables. In addition to N<sub>TFD</sub>, we also obtained ectomycorrhizal community thresholds for nitrogen open field deposition, because open field deposition measurements better reflect the data that is available in spatially mapped deposition datasets<sup>68,97</sup>.

G-tests were performed to test whether host species or soil type influence hyphal and rhizomorph presence or absence. We used logistic regression with each key environmental variable and the presence or absence of emanating hyphae and rhizomorphs within individual OTUs to test for environmental influences on their morphological plasticity. We considered OTUs for which the indicator analysis suggested a response to a particular environmental variable and, for statistical power, we only tested OTUs with  $\geq 15\%$  presence and  $\geq 15\%$  absence of emanating hyphae or rhizomorphs (Extended Data Table 1). Target tree species and soil type were used as co-variables, to account for potential variation in hyphal and rhizomorph development in mycorrhizas belonging to the same OTU among different tree species and different soil types.

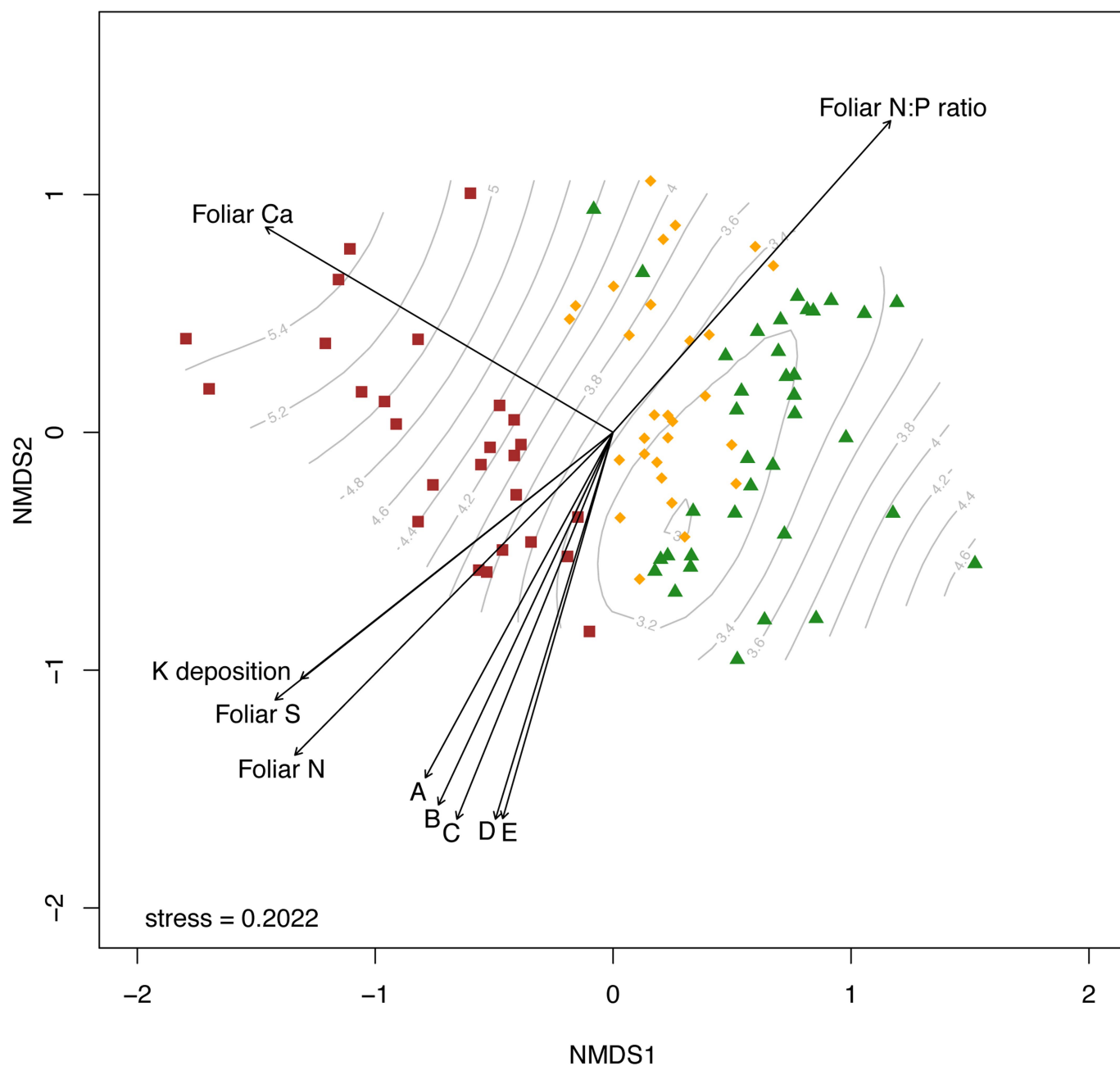
**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.



**Code availability.** R scripts for data analyses are available from the corresponding author upon reasonable request.

**Data availability.** Sequencing data generated during the current study are available through DRYAD (<https://datadryad.org/>) with the DOI <https://doi.org/10.5061/dryad.cr70qc8>. Morphological characteristic and host specificity data are available from the corresponding author upon reasonable request. All environmental data (including deposition, foliar chemistry, soil and meteorological data) are available from UNECE ICP Forests (<http://icp-forests.net/page/data-requests>). Restrictions apply to the availability of these data, which were used under license for the current study. Data are available from the corresponding author upon reasonable request and with permission of UNECE ICP Forests.

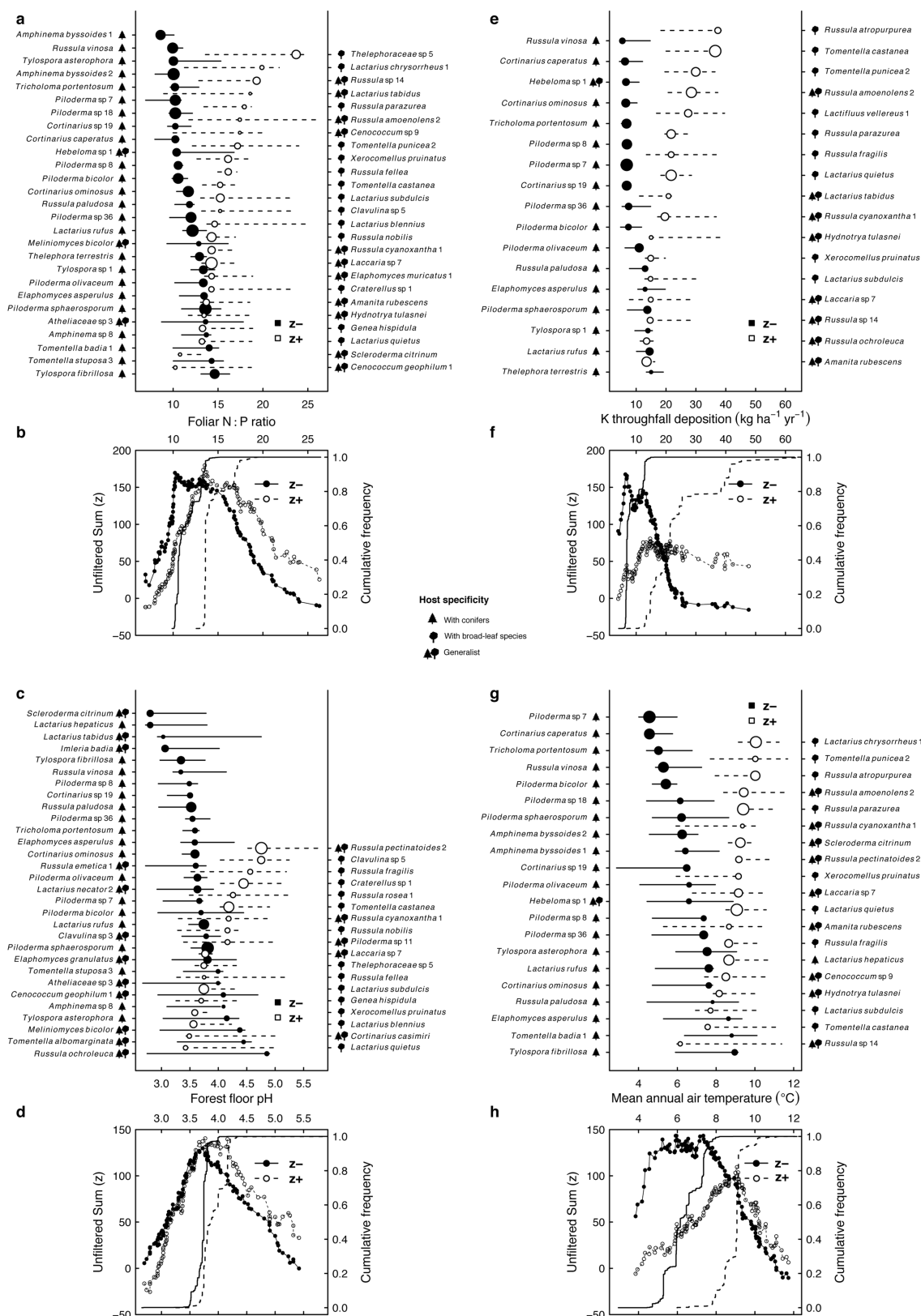
66. Ferretti, M. & Fischer, R. (eds) *Forest Monitoring: Methods for Terrestrial Investigations in Europe with an Overview of North America and Asia Forest Monitoring (Developments in Environmental Science Vol. 12)* (Elsevier, Amsterdam, 2013).
67. de Vries, W. et al. Intensive monitoring of forest ecosystems in Europe: 1. Objectives, set-up and evaluation strategy. *For. Ecol. Manage.* **174**, 77–95 (2003).
68. Dirnböck, T. et al. Forest floor vegetation response to nitrogen deposition in Europe. *Glob. Change Biol.* **20**, 429–440 (2014).
69. MCPFE Liaison Unit Warsaw, UNECE & FAO. *State of Europe's Forests: the MCPFE Report on Sustainable Forest Management in Europe* (Ministerial Conference on the Protection of Forests in Europe, Warsaw, 2007).
70. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2**, 113–118 (1993).
71. White, T. J., Bruns, T., Lee, S. & Taylor, J. in *PCR Protocols: a Guide to Methods and Applications* (eds Innis, M. A. et al.) 315–322 (Academic, New York, 1990).
72. UNECE ICP Forests Programme Co-ordinating Centre (ed.). *Manual on Methods and Criteria for Harmonized Sampling, Assessment, Monitoring and Analysis of the Effects of Air Pollution on Forests* (Thünen Institute for Forest Ecosystems, Eberswalde, 2016).
73. IUSS Working Group WRB. *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps (World Soil Resources Reports 106)* (FAO, Rome, 2015).
74. Eichhorn, J. et al. in *Manual on Methods and Criteria for Harmonized Sampling, Assessment, Monitoring and Analysis of the Effects of Air Pollution on Forests* (ed. UNECE ICP Forests Programme Co-ordinating Centre) 54 (Thünen Institute for Forest Ecosystems, Eberswalde, 2016).
75. Rautio, P., Fürst, A., Stefan, K. & Bartels, U. in *Manual on Methods and Criteria for Harmonized Sampling, Assessment, Monitoring and Analysis of the Effects of Air Pollution on Forests* (ed. UNECE ICP Forests Programme Co-ordinating Centre) 19 (Thünen Institute for Forest Ecosystems, Eberswalde, 2016).
76. Waldner, P. et al. Detection of temporal trends in atmospheric deposition of inorganic nitrogen and sulphate to forests in Europe. *Atmos. Environ.* **95**, 363–374 (2014).
77. Raspe, S., Beuker, E., Preuhsler, T. & Bastrup-Birk, A. in *Manual on Methods and Criteria for Harmonized Sampling, Assessment, Monitoring and Analysis of the Effects of Air Pollution on Forests* (ed. UNECE ICP Forests Programme Co-ordinating Centre) 35 (Thünen Institute for Forest Ecosystems, Eberswalde, 2016).
78. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
79. Kears, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
80. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
81. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
82. Kõljalg, U. et al. Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22**, 5271–5277 (2013).
83. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
84. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
85. Rinaldi, A. C., Comandini, O. & Kuyper, T. W. Ectomycorrhizal fungal diversity: separating the wheat from the chaff. *Fungal Divers.* **33**, 1–45 (2008).
86. Tedersoo, L., May, T. W. & Smith, M. E. Ectomycorrhizal lifestyle in fungi: global diversity, distribution, and evolution of phylogenetic lineages. *Mycorrhiza* **20**, 217–263 (2010).
87. Legendre, P. & Gallagher, E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271–280 (2001).
88. R Core Team. R: A language and environment for statistical computing. <http://www.R-project.org/> (R Foundation for Statistical Computing, Vienna, 2016).
89. Borcard, D., Legendre, P. & Drapeau, P. Partialling out the spatial component of ecological variation. *Ecology* **73**, 1045–1055 (1992).
90. Legendre, P. & Legendre, L. *Numerical Ecology* 2nd edn (Springer, Amsterdam, 1998).
91. Blanchet, F. G., Legendre, P. & Borcard, D. Forward selection of explanatory variables. *Ecology* **89**, 2623–2632 (2008).
92. Nychka, D., Furrer, R., Paige, J. & Sain, S. fields: tools for spatial data. <http://www.image.ucar.edu/fields> (2015).
93. Lee, C.-R. et al. On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nat. Commun.* **8**, 14458 (2017).
94. Lamb, A. M. et al. Climate-driven mitochondrial selection: a test in Australian songbirds. *Mol. Ecol.* **27**, 898–918 (2018).
95. Kalogirou, S. lctools: local correlation, spatial inequalities, geographically weighted regression and other tools. <https://CRAN.R-project.org/package=lctools> (2016).
96. Baker, M. E. & King, R. S. A new method for detecting and interpreting biodiversity and ecological community thresholds. *Methods Ecol. Evol.* **1**, 25–37 (2010).
97. Dore, A. J. et al. Evaluation of the performance of different atmospheric chemical transport models and inter-comparison of nitrogen and sulphur deposition estimates for the UK. *Atmos. Environ.* **119**, 131–143 (2015).



**Extended Data Fig. 1 | Global non-metric multidimensional scaling ordination of community composition.** Plots shown with host trees: brown squares, beech; blue circles, oak; green triangles, pine; yellow diamonds, spruce. Isoclines depict the forest-floor pH and arrows

NMDS1

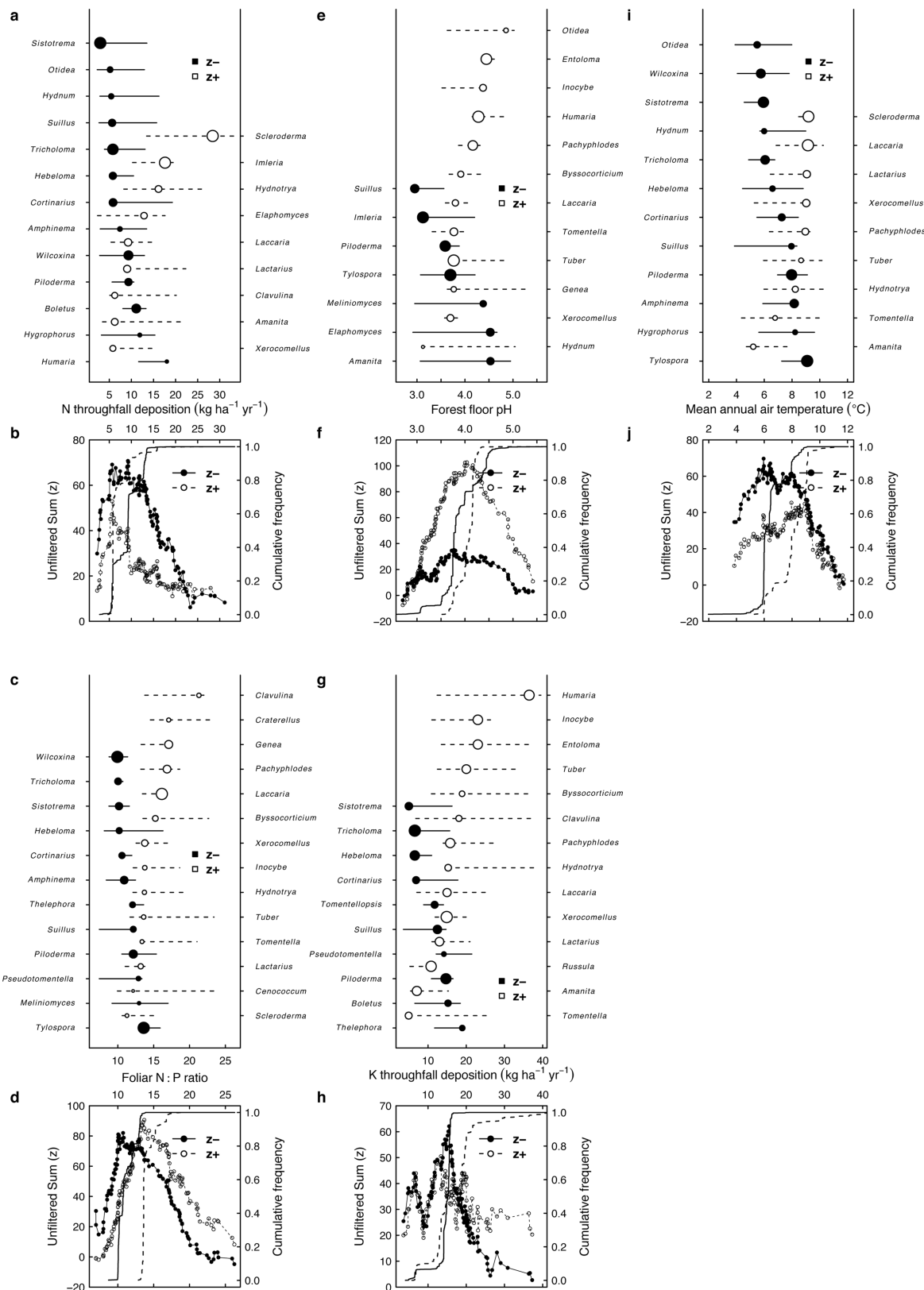
show the direction and strength of correlation of the most-influential environmental variables according to their  $R^2$  values ( $>0.4$ ). A, MAT; B, mean minimum annual air temperature; C, growing season length; D,  $\text{NH}_4$  throughfall deposition; E,  $\text{N}_{\text{TFD}}$ .



**Extended Data Fig. 2 | Threshold indicator taxa analyses.** **a, c, e, g.** Analyses of individual OTU abundances in response to N:P<sub>F</sub> (**a**), forest-floor pH (**c**), K<sub>TFD</sub> (**e**) and MAT (**g**). Black symbols correspond to taxa declining with the increasing variable (z-) and open symbols depict increasing taxa (z+). Symbol size is proportional to magnitude of response (z-score). Horizontal lines represent 5th and 95th quantiles

of values resulting in the largest change in taxon z-scores among 1,000 bootstrap replicates. Tree shapes indicate host generalist, conifer- or broad-leaf-specific. **b, d, f, h.** Community-level output of accumulated z-scores per plot is shown in response to N:P<sub>F</sub> (**b**), forest-floor pH (**d**), K<sub>TFD</sub> (**f**) and MAT (**h**).





**Extended Data Fig. 3 | Threshold indicator taxa analysis at the genus level.** **a, c, e, g, i,** Analyses in response to  $N_{\text{TDF}}$  (**a**),  $N:P_F$  (**c**), forest-floor pH (**e**),  $K_{\text{TDF}}$  (**g**) and MAT (**i**). Black symbols correspond to taxa that declined with the increasing variable (z-) and open symbols depict increasing taxa (z+). Symbol size is proportional to magnitude of response

(z-score). Horizontal lines represent 5th and 95th quantiles of values resulting in the largest change in taxon z-scores among 1,000 bootstrap replicates. **b, d, f, h, j,** The community-level output of the accumulated z-scores per plot is shown in response to  $N_{\text{TDF}}$  (**b**),  $N:P_F$  (**d**), forest-floor pH (**f**),  $K_{\text{TDF}}$  (**h**) and MAT (**j**).

Extended Data Table 1 | Envfit results for the environmental variables used in the non-metric multi-dimensional scaling ordination

	NMDS1	NMDS2	$R^2$	$P$ -value
VECTORS:				
<b>Latitude</b>	<b>0.95129</b>	<b>0.3083</b>	<b>0.2462</b>	<b>0.001</b>
<b>Longitude</b>	<b>0.22325</b>	<b>0.97476</b>	<b>0.3762</b>	<b>0.001</b>
Defoliation	-0.56313	-0.82637	0.0178	0.379
<b>Foliar N</b>	<b>-0.71074</b>	<b>-0.70345</b>	<b>0.5719</b>	<b>0.001</b>
<b>Foliar K</b>	<b>-0.77982</b>	<b>-0.626</b>	<b>0.3201</b>	<b>0.001</b>
<b>Dep tf pH</b>	<b>-0.98433</b>	<b>-0.17634</b>	<b>0.3938</b>	<b>0.001</b>
<b>MAT</b>	<b>-0.48638</b>	<b>-0.87375</b>	<b>0.4263</b>	<b>0.001</b>
<b>Min annual air temperature</b>	<b>-0.43426</b>	<b>-0.90079</b>	<b>0.4749</b>	<b>0.001</b>
<b>Max annual air temperature</b>	<b>-0.35748</b>	<b>-0.93392</b>	<b>0.1344</b>	<b>0.002</b>
<b>Growth season length</b>	<b>-0.38123</b>	<b>-0.92448</b>	<b>0.4845</b>	<b>0.001</b>
<b>Soil pH CaCl<sub>2</sub></b>	<b>-0.37544</b>	<b>0.92685</b>	<b>0.322</b>	<b>0.001</b>
<b>Forest floor pH CaCl<sub>2</sub></b>	<b>-0.85397</b>	<b>0.52032</b>	<b>0.4833</b>	<b>0.001</b>
Forest floor total organic C	0.17461	0.98464	0.02	0.366
<b>Forest floor total organic N</b>	<b>-0.9989</b>	<b>-0.04683</b>	<b>0.1181</b>	<b>0.003</b>
<b>Elevation</b>	<b>-0.40274</b>	<b>0.91532</b>	<b>0.2059</b>	<b>0.001</b>
Foliar P	0.48318	0.87552	0.0265	0.234
<b>Foliar N:P ratio</b>	<b>0.67447</b>	<b>0.7383</b>	<b>0.4888</b>	<b>0.001</b>
<b>Foliar Ca</b>	<b>-0.86227</b>	<b>0.50645</b>	<b>0.4524</b>	<b>0.001</b>
<b>Foliar Mg</b>	<b>0.89284</b>	<b>-0.45037</b>	<b>0.3697</b>	<b>0.001</b>
<b>Foliar S</b>	<b>-0.80043</b>	<b>-0.59943</b>	<b>0.511</b>	<b>0.001</b>
<b>Dep tf conductivity</b>	<b>-0.31098</b>	<b>-0.95042</b>	<b>0.3403</b>	<b>0.001</b>
<b>Dep tf K</b>	<b>-0.79468</b>	<b>-0.60703</b>	<b>0.4463</b>	<b>0.001</b>
<b>Dep tf Ca</b>	<b>-0.99963</b>	<b>-0.02728</b>	<b>0.165</b>	<b>0.001</b>
<b>Dep tf Mg</b>	<b>-0.63371</b>	<b>-0.77357</b>	<b>0.1937</b>	<b>0.001</b>
<b>Dep tf Na</b>	<b>0.27266</b>	<b>0.96211</b>	<b>0.3241</b>	<b>0.001</b>
<b>Dep tf NH<sub>4</sub></b>	<b>-0.29206</b>	<b>-0.9564</b>	<b>0.4298</b>	<b>0.001</b>
<b>Dep tf Cl</b>	<b>0.26542</b>	<b>0.96413</b>	<b>0.3687</b>	<b>0.001</b>
<b>Dep tf NO<sub>3</sub></b>	<b>-0.36813</b>	<b>-0.92977</b>	<b>0.2465</b>	<b>0.001</b>
<b>Dep tf SO<sub>4</sub></b>	<b>0.29501</b>	<b>0.95549</b>	<b>0.2802</b>	<b>0.001</b>
<b>Dep tf N total</b>	<b>-0.30962</b>	<b>-0.95086</b>	<b>0.4363</b>	<b>0.001</b>
Mean annual precipitation	0.98576	0.16817	0.0484	0.074
Soil total organic C	-0.46567	-0.88496	0.0507	0.07
<b>Soil total organic N</b>	<b>-0.87811</b>	<b>-0.47846</b>	<b>0.0963</b>	<b>0.007</b>
<b>Forest floor dry mass</b>	<b>0.78238</b>	<b>-0.62281</b>	<b>0.1256</b>	<b>0.001</b>
<b>Mean tree age</b>	<b>-0.93483</b>	<b>0.35511</b>	<b>0.1855</b>	<b>0.001</b>
FACTORS:				
<b>Tree species</b>			<b>0.5262</b>	<b>0.001</b>
<b>Soil type</b>			<b>0.2556</b>	<b>0.001</b>
<b>Climatic region</b>			<b>0.3125</b>	<b>0.001</b>

Significant variables are shown in bold.

**Extended Data Table 2 | Observed and expected frequencies of hyphae and rhizomorph presence**

**a**

	Hyphal presence			Rhizomorph presence		
	beech	pine	spruce	beech	pine	spruce
Observed	4451	3767	4709	195	443	258
Expected	4525	3777	4636	313	262	321

**b**

	Hyphal presence						Rhizomorph presence					
	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6
Observed	4412	1067	6064	232	622	490	363	47	373	42	31	40
Expected	3978	1159	6204	258	644	644	277	81	431	18	45	45

**a.** Hyphal and rhizomorph presence per host tree species. **b.** Hyphal and rhizomorph presence per soil type. S1, Fe–Al soils; S2, clay soils; S3, soils with little or no differentiation; S4, salt accumulation soils; S5, organic accumulation soils; S6, limited root soil.



Extended Data Table 3 | Effects of key variables on hyphal plasticity

a						
OTU	n	N <sub>TFD</sub>	N:P <sub>F</sub>	Forest floor pH	MAT	K <sub>TFD</sub>
<i>Amphinema byssoides</i> 1	129		0.06 <sup>-</sup>		0.19 <sup>-</sup>	
<i>Atheliaceae</i> sp 3	72		0.61 <sup>-</sup>	0.35 <sup>-</sup>		
<i>Cortinarius caperatus</i>	55	0.64 <sup>-</sup>	↓ <0.05 <sup>-</sup>		0.061 <sup>-</sup>	0.061 <sup>-</sup>
<i>Elaphomyces asperulus</i>	347	0.52 <sup>-</sup>	0.24 <sup>-</sup>	0.62 <sup>-</sup>	0.16 <sup>-</sup>	0.27 <sup>-</sup>
<i>Elaphomyces granulatus</i>	519	↑ <0.0001 <sup>+</sup>		↓ <0.001 <sup>-</sup>		
<i>Elaphomyces muricatus</i> 1	438	0.14 <sup>+</sup>	0.57 <sup>+</sup>			
<i>Imleria badia</i>	115	0.32 <sup>+</sup>		0.74 <sup>-</sup>		
<i>Laccaria</i> sp 7	138	0.11 <sup>+</sup>	0.67 <sup>+</sup>	↑ <0.05 <sup>+</sup>	0.94 <sup>+</sup>	0.8 <sup>+</sup>
<i>Lactarius blennius</i>	159		0.94 <sup>+</sup>	0.25 <sup>+</sup>		
<i>Lactarius hepaticus</i>	98	↑ <0.01 <sup>+</sup>		↓ <0.01 <sup>-</sup>	↓ <0.0001 <sup>+</sup>	
<i>Lactarius subdulcis</i>		↓ <0.05 <sup>+</sup>	↓ <0.01 <sup>+</sup>	0.67 <sup>+</sup>		
<i>Meliniomyces bicolor</i>	108	0.38 <sup>-</sup>	0.59 <sup>-</sup>	0.96 <sup>-</sup>	0.39 <sup>-</sup>	
<i>Piloderma bicolor</i>	102	0.27 <sup>-</sup>	0.19 <sup>-</sup>	0.19 <sup>-</sup>	0.75 <sup>-</sup>	0.23 <sup>-</sup>
<i>Piloderma</i> sp 7	56	↑ <0.05 <sup>-</sup>	0.054 <sup>-</sup>	0.23 <sup>-</sup>	↑ <0.01 <sup>-</sup>	↑ <0.01 <sup>-</sup>
<i>P. sphaerosporum</i>	1051	↑ <0.0001 <sup>-</sup>	↑ <0.0001 <sup>-</sup>	↑ <0.001 <sup>-</sup>	↑ <0.0001 <sup>-</sup>	↑ <0.0001 <sup>-</sup>
<i>Russula cyanoxantha</i> 1	78	0.086 <sup>+</sup>	↓ <0.05 <sup>+</sup>	0.28 <sup>+</sup>	0.26 <sup>+</sup>	0.09 <sup>+</sup>
<i>Russula ochroleuca</i>	834	↑ <0.01 <sup>+</sup>		0.12 <sup>-</sup>		0.96 <sup>+</sup>
<i>Russula</i> sp 14	225	0.078 <sup>+</sup>	↓ <0.001 <sup>+</sup>		0.12 <sup>+</sup>	↓ <0.05 <sup>+</sup>
<i>Thelephora terrestris</i>	64		0.74 <sup>-</sup>			0.78 <sup>-</sup>
<i>Tomentella albomarginata</i>	100	0.1 <sup>+</sup>		0.075 <sup>-</sup>	0.47 <sup>+</sup>	
<i>Tomentella castanea</i>	78		0.093 <sup>+</sup>	0.22 <sup>+</sup>	0.064 <sup>+</sup>	0.34 <sup>+</sup>
<i>Tylospora asterophora</i>	355	↑ <0.001 <sup>-</sup>	↑ <0.05 <sup>-</sup>	0.62 <sup>-</sup>	↑ <0.05 <sup>-</sup>	
<i>Tylospora fibrillosa</i>	555		0.23 <sup>-</sup>	0.28 <sup>-</sup>	0.82 <sup>-</sup>	
<i>Xerocomellus pruinatus</i>	106	0.28 <sup>+</sup>	0.51 <sup>+</sup>	0.08 <sup>+</sup>	0.23 <sup>+</sup>	0.48 <sup>+</sup>

b						
OTU	n	N <sub>TFD</sub>	N:P <sub>F</sub>	Forest floor pH	MAT	K <sub>TFD</sub>
<i>Amphinema byssoides</i> 1	119		0.086 <sup>-</sup>		0.17 <sup>-</sup>	
<i>Atheliaceae</i> sp 3	72		0.66 <sup>-</sup>	0.22 <sup>-</sup>		
<i>Elaphomyces asperulus</i>	332	0.16 <sup>-</sup>	↑ <0.05 <sup>-</sup>	0.94 <sup>-</sup>	↑ <0.01 <sup>-</sup>	0.20 <sup>-</sup>
<i>Elaphomyces granulatus</i>	538	↑ <0.0001 <sup>+</sup>		↓ <0.0001 <sup>-</sup>		
<i>Elaphomyces muricatus</i> 1	376	0.33 <sup>+</sup>	0.19 <sup>+</sup>			
<i>Imleria badia</i>	86	0.22 <sup>+</sup>		0.89 <sup>-</sup>		
<i>Lactarius blennius</i>	132		0.75 <sup>+</sup>	0.71 <sup>+</sup>		
<i>Lactarius hepaticus</i>	86	↑ <0.01 <sup>+</sup>		↓ <0.05 <sup>-</sup>	↓ <0.0001 <sup>+</sup>	
<i>Lactarius subdulcis</i>	241	↓ <0.05 <sup>+</sup>	↓ <0.01 <sup>+</sup>	0.75 <sup>+</sup>	0.35 <sup>+</sup>	0.94 <sup>+</sup>
<i>Piloderma sphaerosporum</i>	869	↑ <0.0001 <sup>-</sup>	↑ <0.0001 <sup>-</sup>	↑ <0.01 <sup>-</sup>	↑ <0.0001 <sup>-</sup>	↑ <0.0001 <sup>-</sup>
<i>Russula cyanoxantha</i> 1	72	0.48 <sup>+</sup>	0.054 <sup>+</sup>	0.059 <sup>+</sup>	0.21 <sup>+</sup>	0.06 <sup>+</sup>
<i>Russula ochroleuca</i>	861	↑ <0.05 <sup>+</sup>		0.079 <sup>-</sup>		0.86 <sup>+</sup>
<i>Russula</i> sp 14	116	0.87 <sup>+</sup>	↓ <0.01 <sup>+</sup>		0.84 <sup>+</sup>	0.064 <sup>+</sup>
<i>Thelephora terrestris</i>	54		0.86 <sup>-</sup>			0.94 <sup>-</sup>
<i>Tomentella castanea</i>	76		↓ <0.05 <sup>+</sup>	0.076 <sup>+</sup>	0.2 <sup>+</sup>	0.55 <sup>+</sup>
<i>Tylospora asterophora</i>	325	↑ <0.01 <sup>-</sup>	0.081 <sup>-</sup>	0.64 <sup>-</sup>	0.21 <sup>-</sup>	
<i>Tylospora fibrillosa</i>	71		↑ <0.05 <sup>-</sup>	0.27 <sup>-</sup>	0.23 <sup>-</sup>	
<i>Xerocomellus pruinatus</i>	63	0.32 <sup>+</sup>	0.11 <sup>+</sup>	0.58 <sup>+</sup>	0.77 <sup>+</sup>	0.37 <sup>+</sup>

a, OTUs with 97% sequence similarity. b, OTUs with 99% sequence similarity. Values with  $P < 0.05$  are shown in bold. Logistic regressions were calculated only for OTUs for which the indicator analysis suggested a response to a particular environmental variable. Values with superscript - refer to a declining indicator (z-); values with superscript + refer to an increasing indicator (z+). A downward-pointing arrow denotes a negative correlation; an upward-pointing arrow denotes a positive correlation.

Extended Data Table 4 | Effects of key variables on rhizomorph plasticity

<b>a</b>						
OTU	n	N <sub>TFD</sub>	N:P <sub>F</sub>	Forest floor pH	MAT	K <sub>TFD</sub>
<i>Amphinema byssoides</i>	129		0.22 <sup>-</sup>		0.42	
<i>Imleria badia</i>	115	0.87 <sup>+</sup>				
<i>Piloderma bicolor</i>	102	0.28 <sup>-</sup>	0.89 <sup>-</sup>	0.25 <sup>-</sup>	0.25 <sup>-</sup>	0.75 <sup>-</sup>
<i>P. olivaceum</i>	409	↓<0.01 <sup>-</sup>	0.32 <sup>-</sup>	0.45 <sup>-</sup>	0.24 <sup>-</sup>	0.24 <sup>-</sup>
<i>P. sp 36</i>	58	0.073 <sup>-</sup>	↓<0.05 <sup>-</sup>	0.17 <sup>-</sup>	↓<0.05 <sup>-</sup>	0.17 <sup>-</sup>
<i>P. sphaerosporum</i>	1051	0.21 <sup>-</sup>	0.76 <sup>-</sup>	↑<0.05 <sup>-</sup>	0.27 <sup>-</sup>	0.94 <sup>-</sup>
<i>Xerocomellus pruinatus</i>	106	↓<0.01 <sup>+</sup>	0.11 <sup>+</sup>	↓<0.05 <sup>+</sup>	0.60 <sup>+</sup>	↓<0.01 <sup>+</sup>

<b>b</b>						
OTU	n	N <sub>TFD</sub>	N:P <sub>F</sub>	Forest floor pH	MAT	K <sub>TFD</sub>
<i>Amphinema byssoides</i>	119		0.29 <sup>-</sup>		0.28	
<i>Imleria badia</i>	86	0.40 <sup>+</sup>				
<i>P. olivaceum</i>	126	↓<0.05 <sup>-</sup>	0.64 <sup>-</sup>	0.94 <sup>-</sup>	0.051 <sup>-</sup>	0.39 <sup>-</sup>
<i>P. sp 36</i>	58	0.095 <sup>-</sup>	0.093 <sup>-</sup>	0.20 <sup>-</sup>	↓<0.05 <sup>-</sup>	0.24 <sup>-</sup>
<i>P. sphaerosporum</i>	869	0.21 <sup>-</sup>	0.54 <sup>-</sup>	0.081 <sup>-</sup>	0.46 <sup>-</sup>	0.83 <sup>-</sup>
<i>Xerocomellus pruinatus</i>	63	0.058 <sup>+</sup>	0.31 <sup>+</sup>	↓<0.05 <sup>+</sup>	0.77 <sup>+</sup>	↓<0.05 <sup>+</sup>

**a**, OTUs with 97% sequence similarity. **b**, OTUs with 99% sequence similarity. Values with  $P < 0.05$  are shown in bold. Logistic regressions were calculated only for OTUs for which the indicator analysis suggested a response to a particular environmental variable. Values with superscript – refer to a declining indicator ( $z^-$ ); values with superscript + refer to an increasing indicator ( $z^+$ ). A downward-pointing arrow denotes a negative correlation; an upward-pointing arrow denotes a positive correlation.

Extended Data Table 5 | Effects of key variables on hyphal and rhizomorph presence on the total ectomycorrhizal community

	<b>N<sub>TFD</sub></b>	<b>N:P<sub>F</sub></b>	<b>Forest floor pH</b>	<b>MAT</b>	<b>K<sub>TFD</sub></b>
Hyphae	<b>↑&lt;.0001</b>	<b>↓&lt;.0001</b>	<b>↓&lt;.0001</b>	0.26	<b>↓&lt;0.05</b>
Rhizomorphs	<b>↓&lt;.0001</b>	<b>↓&lt;.0001</b>	<b>↓&lt;0.001</b>	<b>↓&lt;.0001</b>	<b>↓&lt;.0001</b>

Values with  $P < 0.05$  are shown in bold. Values with superscript – refer to a declining indicator (z–); values with superscript + refer to an increasing indicator (z+). A downward-pointing arrow denotes a negative correlation; an upward-pointing arrow denotes a positive correlation.



# Visualizing late states of human 40S ribosomal subunit maturation

Michael Ameismeier<sup>1,2</sup>, Jingdong Cheng<sup>1,2</sup>, Otto Berninghausen<sup>1</sup> & Roland Beckmann<sup>1\*</sup>

**The formation of eukaryotic ribosomal subunits extends from the nucleolus to the cytoplasm and entails hundreds of assembly factors. Despite differences in the pathways of ribosome formation, high-resolution structural information has been available only from fungi. Here we present cryo-electron microscopy structures of late-stage human 40S assembly intermediates, representing one state reconstituted in vitro and five native states that range from nuclear to late cytoplasmic. The earliest particles reveal the position of the biogenesis factor RRP12 and distinct immature rRNA conformations that accompany the formation of the 40S subunit head. Molecular models of the late-acting assembly factors TSR1, RIOK1, RIOK2, ENP1, LTV1, PNO1 and NOB1 provide mechanistic details that underlie their contribution to a sequential 40S subunit assembly. The NOB1 architecture displays an inactive nuclease conformation that requires rearrangement of the PNO1-bound 3' rRNA, thereby coordinating the final rRNA folding steps with site 3 cleavage.**

Ribosomes are universally conserved macromolecular complexes that translate mRNA into protein. In eukaryotes, they consist of a small 40S and a large 60S subunit that together comprise four rRNAs and about 80 ribosomal proteins. During their assembly, a multitude of over 200 *trans*-acting ribosome biogenesis factors (RBFs) ensures correct cleavage, modification and folding of rRNA and concomitant incorporation of ribosomal proteins<sup>1–3</sup>.

In eukaryotes, ribosome biogenesis starts in the nucleolus with the transcription of three of the four rRNAs as a primary transcript<sup>4</sup>. After dissociation of the large subunit precursor, pre-40S particles are exported to the cytoplasm, where final maturation processes occur<sup>5</sup>. Many proteins are involved in the later stages of human small subunit maturation, several of which we observe in this work, including the methyltransferase BUD23 (also known as WBSR22) together with TRMT112, the armadillo (ARM)-like protein RRP12, structural proteins ENP1 (also known as BYSL) and LTV1, the endonuclease NOB1 with its binding partner PNO1, GTPase-like protein TSR1, as well as the atypical kinases RIOK1 and RIOK2<sup>6–8</sup>.

High-resolution structures of pre-40S particles from fungi such as *Saccharomyces cerevisiae* have recently been published<sup>9,10</sup>. Despite differences in rRNA processing<sup>11</sup> and the composition and function of RBFs<sup>12,13</sup>, however, human ribosomal precursors have been only described at low resolution so far<sup>14</sup>, limiting the information to overall positioning of RBFs.

Here we report cryo-electron microscopy (cryo-EM) reconstructions of several small ribosomal subunit precursors that provide detailed insights into late human 40S maturation principles. As a reoccurring concept, we found that assembly factors stabilize rRNA segments in distinct immature conformations before allowing their accommodation into their respective mature positions. Thereby, accurate step-wise and sequential maturation of rRNA from 5' to 3' is ensured, and premature engagement of 40S precursors in 80S formation and erroneous translation is prevented.

## Cryo-EM analysis of the human pre-40S ribosome

To understand better the maturation process of human 40S subunits, we aimed at structurally analysing native complexes, purified with PNO1

as bait (states A–E), and reconstituted pre-40S particles in vitro (state R). Single particle cryo-EM analysis (Extended Data Fig. 1) resulted in a series of pre-40S structures that represent six different middle to late assembly states (Fig. 1a and Extended Data Fig. 2a, b).

In state A, we found the RBFs RRP12, ENP1 and PNO1 together with the methyltransferases BUD23 and TRMT112. State B lacks BUD23 and TRMT112, but comprises LTV1 and ribosomal proteins uS2, uS5 and eS21, as well as the endonuclease NOB1. From state C on, RRP12 is dissociated, and its place is occupied by RACK1. In addition, the kinase RIOK2, and ribosomal proteins eS12 and eS31 have bound, revealing remarkable conservation when compared with highly similar pre-40S subunits from yeast<sup>9,10</sup> (Extended Data Fig. 2d). In state D, ENP1 and LTV1 have been replaced by ribosomal proteins uS3, uS10, eS10 and uS14. Finally, an unassigned protein (factor X) is present in state E, partially occupying the binding site of the C-terminal helix of RIOK2.

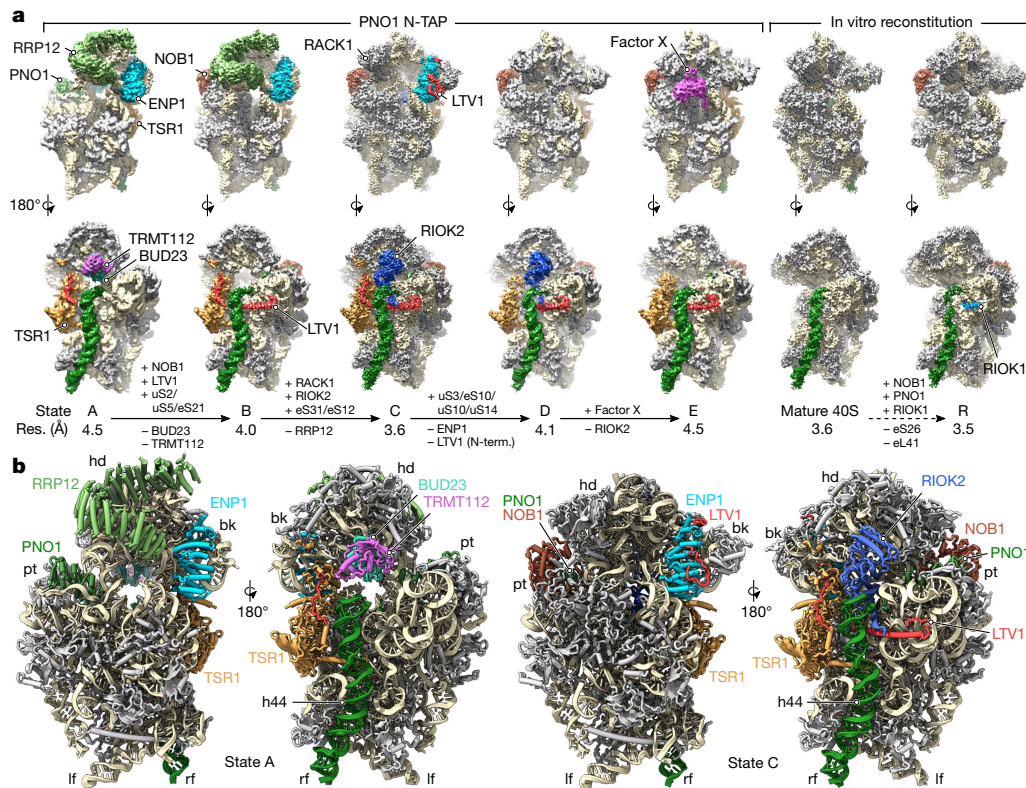
The atypical kinase RIOK1 has also been described to have a role in the final steps of pre-40S maturation, together with PNO1 and NOB1<sup>7</sup>. We did not observe such a state in our native pull-outs and, therefore, attempted to reconstitute it in vitro using purified mature 40S subunits and recombinant proteins (state R). We found that PNO1 and RIOK1 have displaced ribosomal proteins eS26 and eL41, and that NOB1, PNO1 and the pre-18S rRNA in state R adopt a highly similar conformation as in state E, including an immature helix 44 (h44) and a shifted 3' end.

Among all pre-40S reconstructions, two (states C and R) could be refined to average resolutions below 4 Å. A full model of state C was built (Fig. 1b and Extended Data Fig. 2c), including biogenesis factors TSR1, NOB1, PNO1, ENP1 and RIOK2, as well as parts of LTV1, and a short segment of RIOK1 was built in state R (Extended Data Fig. 2e). The average resolution of the other states was between 4 and 4.5 Å, with considerably lower local resolution in flexible areas. It was, however, sufficient to allow for unambiguous rigid body fitting of our models and assignment of RRP12, BUD23 and TRMT112 to extra densities in states A and B (Fig. 1b).

## The maturation process of the 3' major domain

The maturation process of the central region around helices h35–h40 is dominated by large shifts of rRNA, in which RRP12 has a crucial

<sup>1</sup>Gene Center Munich and Center of Integrated Protein Science-Munich (CiPS-M), Department of Biochemistry, University of Munich, Munich, Germany. <sup>2</sup>These authors contributed equally: Michael Ameismeier, Jingdong Cheng. \*e-mail: [beckmann@genzentrum.lmu.de](mailto:beckmann@genzentrum.lmu.de)



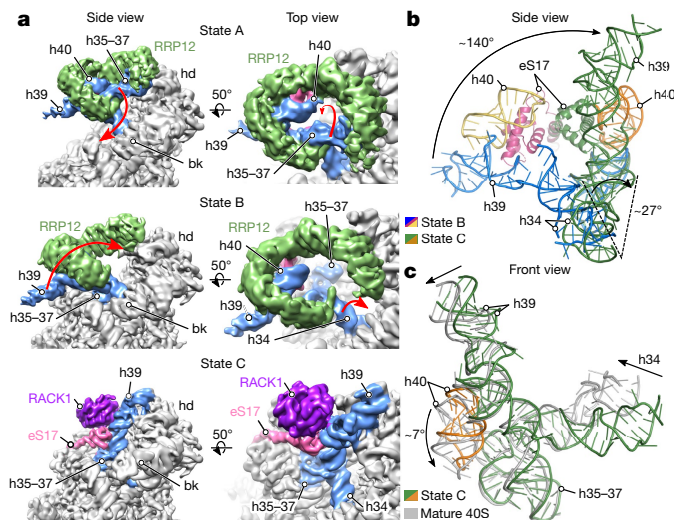
**Fig. 1 | Structural analysis of human pre-40S particles.** **a**, Cryo-EM reconstructions of five native pre-40S particles, obtained via N-terminal tandem affinity purification (N-TAP), and the mature 40S with a precursor-like complex reconstituted in vitro (see Methods). Names,

average resolution and changes in protein composition are stated below. **b**, Models of state A (left) and state C (right) 40S precursors with ribosomal biogenesis factors and h44 highlighted (grey: ribosomal proteins, light brown: rRNA). bk, beak; hd, head; lf, left foot; pt, platform; rf, right foot.

role (Fig. 2 and Extended Data Fig. 3a). In state A, it coordinates h35–h37 and the region around the three-way-junction of h38–h39–h40 in their immature position, and with h41 and uS9 parts of the head (Fig. 2a, top). The flipping of helices h35–h37 downwards by about 90° then allows binding of the uS2–uS5–eS21 cluster together with NOB1 (Fig. 2a, middle and Extended Data Fig. 3b). Once RRP12 is released, h39 rotates almost 140° around the h34 axis to interact with h41, while h40 and eS17 accommodate close to their mature position in state C

(Fig. 2a, bottom, and Fig. 2b). With the movement of these central rRNA helices, the interface is then formed for binding of RACK1. In yeast, Rrp12 has been shown to facilitate export of pre-40S particles through direct interaction with nucleoporins and the small GTPase Gsp1p<sup>15</sup>. In combination with our results, this suggests that RRP12 could serve in a quality control step, in which correct head formation and subsequent export are coupled.

Despite major rRNA rearrangements, helices h34, h39 and h40 are yet to settle into their mature position (Fig. 2c), which depends on maturation events within the beak (Extended Data Fig. 3c). In our earlier particles, ENP1 and LTV1 occupy the central position on the beak, preventing the accommodation of the uS3–uS10–eS10–uS14 cluster and the mature positioning of h34, similar to what has been observed in yeast<sup>9</sup>. After the release of ENP1 and LTV1 from the beak in state D, ribosomal proteins uS3, uS10, eS10 and uS14 settle into their respective sites, and this coincides with a movement of h34 closer to its mature position (Extended Data Fig. 3d).



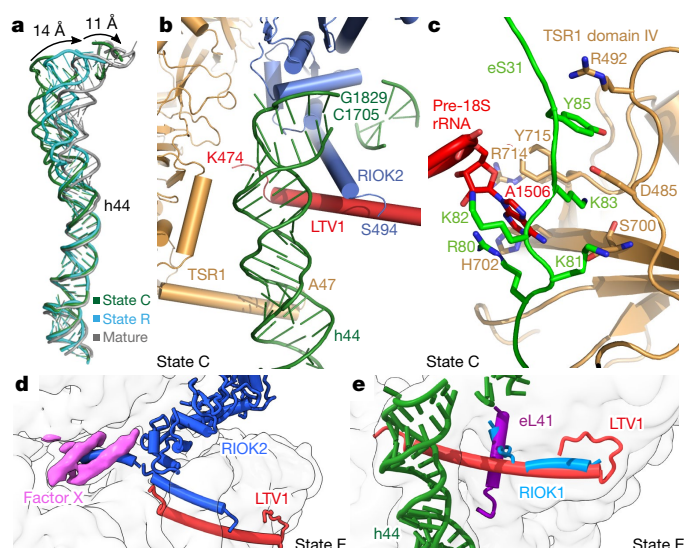
**Fig. 2 | RRP12 binding and sequential pre-18S rRNA rearrangements during head formation.** **a**, Movement of h35–h37 and h39 during early head formation. Volumes filtered at 6 Å. **b**, **c**, Detailed view on rRNA rearrangements. **b**, The transition from state B to C is described by a 140° movement of h39 and h40. **c**, Final maturation requires repositioning of h34, h39 and h40.

### The maturation process of the decoding centre

We observed an immature decoding centre in all native complexes and, to our surprise, also in native 40S particles treated with PNO1, NOB1 and RIOK1 at high molar excess (20×; Extended Data Fig. 4a). Similar to yeast<sup>9</sup>, h44 is shifted outward by approximately 25 Å in states A–E and 11 Å in state R (Fig. 3a). This displacement prevents correct folding of the connecting region between h44 and h28 or h45, and therefore formation of the A and P sites. The biogenesis factors TSR1, LTV1 and RIOK2 probe this immature conformation through distinct terminal helices that bind underneath h44, probably contributing to its stabilization (Fig. 3b).

TSR1, an inactive mimic of translational GTPases<sup>16</sup>, interacts with the pre-40S body near rRNA helices h3–h5, h15 and h44 via domains I and III, with its N-terminal helix passing beneath h44 up to h12 (Fig. 3b and Extended Data Fig. 4b), as well as with the head via its domain





**Fig. 3 | Positioning of h44 and maturation of the decoding centre.**

**a**, Comparison of h44 in its immature and mature position. **b**, Location of terminal helices of TSR1, LTV1 and RIOK2 between displaced h44 and the body. **c**, Interactions between TSR1 and h34 with the N-terminal extension of eS31. **d**, **e**, Clashes of biogenesis factor RIOK2 with unassigned factor X (**d**), RIOK1 with the C terminus of LTV1, and RIOK1 with eL41 (**e**).

IV, as observed in yeast<sup>9</sup>. In addition, we observe distinct interactions between TSR1, pre-18S rRNA and eS31, which is recruited in state C. Together with domain IV of TSR1, the shifted N terminus of eS31 (Extended Data Fig. 4c) coordinates a flipped out A1506 of the junction region of h32, h33 and h34, stabilizing h34 in an immature state after the rearrangement of the h34–h40 region (Fig. 3c). Supported by an insertion in domain II, which binds to uS13 and uS19 in the 3' major domain, domain IV and eS31 might therefore serve as a link between maturation events within the beak and the intersubunit side.

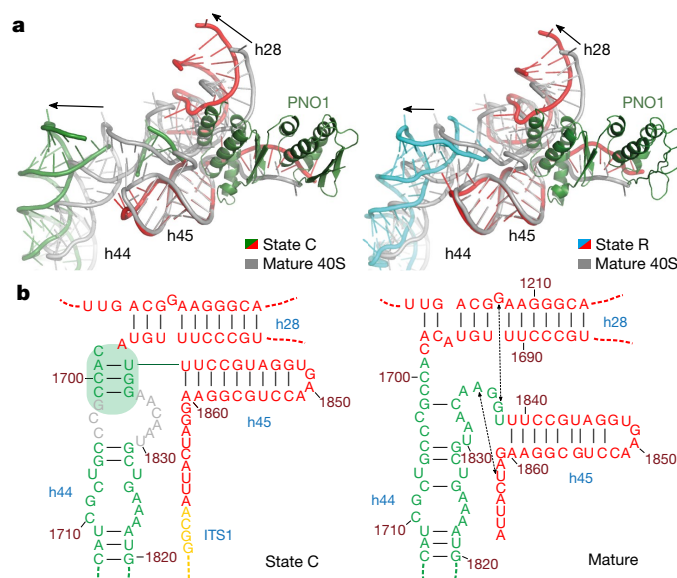
Another connection could represent LTV1, of which we observed two interacting segments (Fig. 1a). In addition to the N-terminal part that binds ENP1 similarly to published structures<sup>17</sup>, its C terminus forms a long helix that stretches for about 60 Å across the intersubunit side and ends beneath h44 (Fig. 3b), a position largely occupied by Dim1 in late yeast 40S precursors<sup>18</sup>. Unlike in yeast, however, human dimethylation by DIMT1 is nuclear and, therefore, DIMT1 is not involved in the final maturation steps in the cytoplasm<sup>8</sup>.

The kinase RIOK2 is observed in states C and D, binding to the decoding centre between the head and the platform as previously described<sup>9</sup>. In addition, we found its C terminus forming an elongated helix, which passes beneath h44 and is deeply buried within the tunnel that connects the solvent and intersubunit side (Fig. 3b). Surprisingly, in state E, factor X occupies parts of this tunnel with two helices, and superimposition of both states reveals overlapping binding sites (Fig. 3d).

Apart from RIOK2, BUD23–TRMT112 and the kinase domain of RIOK1 also bind to the neck and platform of pre-40S particles (Extended Data Fig. 4a, d). Furthermore, the C terminus of RIOK1 binds between h27, h44 and h45, a site previously occupied by LTV1 and later by eL41 in mature 40S particles (Fig. 3e). This is consistent with data suggesting that the C terminus of Rio1 is crucial for its binding to the ribosome in yeast<sup>19</sup>. Furthermore, the clash of RIOK1 with LTV1 and RIOK2 could explain why the sole presence of RIOK1 and not its ATPase activity is sufficient for their release<sup>7</sup>.

### Coordination of site 3 cleavage by PNO1 and NOB1

PNO1 has a decisive role in selectively binding and stabilizing h28 and h44 in an immature conformation in all states (Fig. 4 and Extended Data Fig. 5a). Superimposition of state C with mature 40S reveals two major clashes: First, the last helix of the K homology (KH) 2 domain of PNO1 would clash with h28 (Fig. 4a and Extended Data Fig. 5b, c). This retains



**Fig. 4 | PNO1 interaction determines state of the decoding centre.**

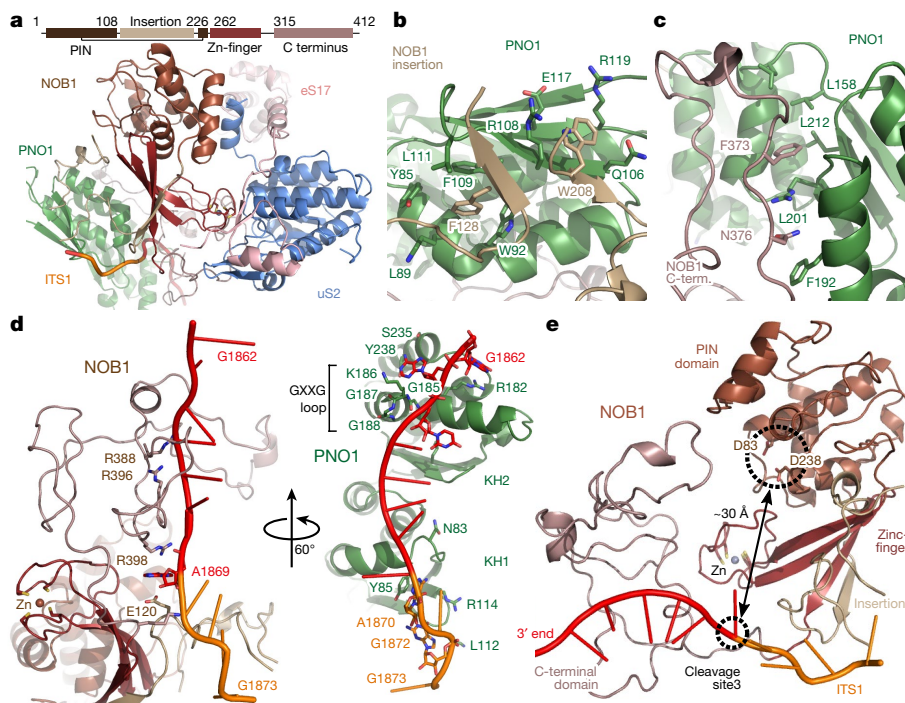
**a**, Conformation of helices h28, h44 and h45 in native state C (left) and the reconstituted particle (state R, right) compared to their mature position (grey). Displacements stabilized by the presence of PNO1 are indicated (arrows). **b**, Immature base pairing as a result of shifted helices h28 and h44. Summary of base interactions in state C and the mature 40S, with a focus on h28, h44 and h45. Base pairings are indicated by black lines and the relevant base stacking interactions by dashed lines. Immature helix highlighted in green. For an extended view, see Extended Data Fig. 6. ITS1, internal transcribed spacer.

h28 in a backward position and prevents stacking between G1207 and G1837 (Extended Data Fig. 5d). Second, PNO1 would also clash with the connecting region between h44 and h45, prohibiting its mature fold while breaking another stacking between A1835 and A1863. As a result, A1863 is flipped out and can be recognized and bound by PNO1 (Extended Data Fig. 5e). Concomitant with the displacement of h44, novel immature base pairing is formed between nucleotides 1699–1701 and 1836–1838 (Fig. 4b and Extended Data Fig. 5f). Notably, as observed in state R, incubation of fully matured 40S subunits with recombinant PNO1 (together with NOB1 and RIOK1) was sufficient to reverse the final maturation step by dissociating eS26 and forcing the rRNA to unfold partially (Fig. 4a, right, and Extended Data Fig. 5b).

Besides stabilizing the central rRNA region in a distinct immature state and consistent with previous results<sup>20</sup>, we found PNO1 directly interacting with NOB1, which is absent in early pre-ribosomal particles<sup>6</sup>. Structurally, NOB1 can be divided into four parts: the PiLT N-terminus (PIN) endonuclease domain, the insertion domain, the zinc-finger domain and the C-terminal domain. The PIN and zinc-finger domains form a stable core of NOB1 (Extended Data Fig. 5g) and mediate its interaction with ribosomal proteins uS2 and eS17 (Fig. 5a). The insertion domain binds to the KH1 domain of PNO1 involving residues Phe128 and Trp208 (Fig. 5b), and the C-terminal part of NOB1 forms a large loop, which interacts with the hydrophobic groove between the two KH domains of PNO1 (Fig. 5c).

In contrast to state R, we can trace several bases into the density for the internal transcribed spacer 1 (ITS1) in all native states (Extended Data Fig. 5h), which is a clear indication that site 3 cleavage has yet to happen. When compared with mature 40S, the 3' end appears shifted from its mature position and is bound by PNO1 and NOB1. The C-terminal domain of NOB1 contains several positively charged residues that stabilize the backbone of the 3' end. In addition, Glu120 of the insertion and Arg398 of the C-terminal extension hold the ultimate nucleotide before site 3, A1869, in a tight pocket. In parallel, PNO1 binds the 3' end and the ITS1 region via residues within its two KH domains, with Arg182, Ser235 and Tyr238 forming hydrogen bonds with G1862 and A1863, contributing to a sequence





**Fig. 5 | NOB1 and PNO1 coordinate site 3 in an inactive conformation.** **a**, Domain architecture and overall structure of NOB1 shows binding of NOB1 by PNO1, uS2 and eS17. **b**, **c**, Molecular detail of the interface of the insertion domain (**b**) and the C-terminal domain (**c**) of NOB1 with

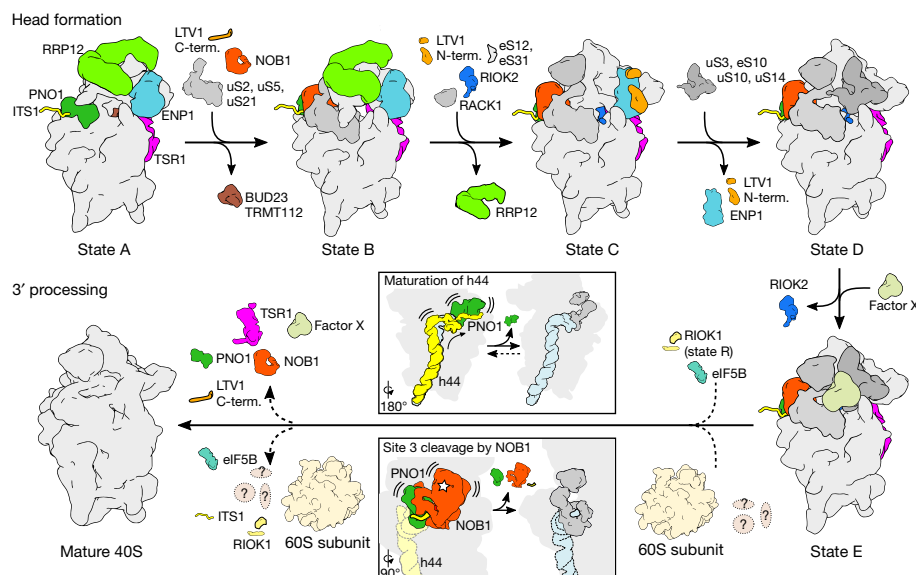
PNO1. **d**, Coordination of the pre-18S rRNA 3' end by NOB1 and PNO1. **e**, Insertion and C terminus of NOB1 shield the site 3 cleavage site from its catalytic centre (dashed circles).

specificity in binding (Fig. 5d and Extended Data Fig. 5h), essentially resembling the situation observed in yeast<sup>9</sup>. Importantly, the RNA around cleavage site 3 is captured at a large distance (approximately 30 Å) from the rigid catalytic centre of the nuclease NOB1 by both, PNO1 and NOB1 itself (Fig. 5e). Movement of the ITS1 towards the catalytic centre is prevented by the insertion and extension domain of NOB1. As a consequence, the PNO1–NOB1 complex is kept in a catalytically silent conformation along the observed maturation process. Clearly, a large conformational change, and probably release of PNO1, will be required to trigger the cleavage of site 3 by releasing

the RNA substrate from its confinement. This would be in agreement with previous hypotheses suggesting that, in the presence of Rio1, cleavage at site D in yeast (site 3 in humans) by Nob1 is facilitated after Pno1 dissociation<sup>21</sup>.

## Conclusion

Taken together, late maturation of human small ribosomal subunit occurs in a sequential manner (Fig. 6): The largely mature 40S body (state A) is further completed by the incorporation of several ribosomal proteins, allowing for recruitment of NOB1 as observed in state



**Fig. 6 | A model of stepwise maturation of the human 40S ribosomal subunit.** Illustrations of distinct human 40S precursors during head formation and 3' processing. Transitions in RBF composition not shown in this work are displayed with dotted arrows. Movement of h44 and cleavage

of site 3 by NOB1 is shown in inserts with focus on PNO1 (green), h44 (yellow and blue), NOB1 (orange) and uS26 and uS28 (grey). The 60S model is based on EMDB-5592.

B. At the same time, the head and beak are formed by going through a sequence of distinct immature rRNA conformations, which is guided by interactions with BUD23, TRMT112, RRP12, ENP1, LTV1, TSR1 and RIOK2, and eventually results in export competence. Once in the cytoplasm, several factors have been suggested to have a role in eukaryotes in triggering the final maturation (adopting native conformation, factor release, site 3 cleavage), including RIOK1, eIF5B (Fun12 in yeast) and mature 60S subunits<sup>22</sup>. Interestingly, eIF5B as well as mature 60S subunits would clash at several sites with our state E, which may contribute to driving these last structural rearrangements (Extended Data Fig. 7). Finally, formation of the decoding region is inevitably accompanied by dissociation or repositioning of PNO1, which, in turn, may facilitate the cleavage of 18S-E rRNA at site 3 and thus concludes 40S formation.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0193-0>.

Received: 22 December 2017; Accepted: 20 April 2018;  
Published online: 06 June 2018

1. Woolford, J. L. Jr & Baserga, S. J. Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics* **195**, 643–681 (2013).
2. Zemp, I. & Kutay, U. Nuclear export and cytoplasmic maturation of ribosomal subunits. *FEBS Lett.* **581**, 2783–2793 (2007).
3. Tschochner, H. & Hurt, E. Pre-ribosomes on the road from the nucleolus to the cytoplasm. *Trends Cell Biol.* **13**, 255–263 (2003).
4. Phipps, K. R., Charette, J. & Baserga, S. J. The small subunit processome in ribosome biogenesis—progress and prospects. *Wiley Interdiscip. Rev. RNA* **2**, 1–21 (2011).
5. Rouquette, J., Choesmel, V. & Gleizes, P. E. Nuclear export and cytoplasmic processing of precursors to the 40S ribosomal subunits in mammalian cells. *EMBO J.* **24**, 2862–2872 (2005).
6. Wyler, E. et al. Tandem affinity purification combined with inducible shRNA expression as a tool to study the maturation of macromolecular assemblies. *RNA* **17**, 189–200 (2011).
7. Widmann, B. et al. The kinase activity of human Rio1 is required for final steps of cytoplasmic maturation of 40S subunits. *Mol. Biol. Cell* **23**, 22–35 (2012).
8. Zorbas, C. et al. The human 18S rRNA base methyltransferases DIMT1L and WBSCR22-TRMT112 but not rRNA modification are required for ribosome biogenesis. *Mol. Biol. Cell* **26**, 2080–2095 (2015).
9. Heuer, A. et al. Cryo-EM structure of a late pre-40S ribosomal subunit from *Saccharomyces cerevisiae*. *eLife* **6**, e30189 (2017).
10. Scaiola, A. et al. Structure of a eukaryotic cytoplasmic pre-40S ribosomal subunit. *EMBO J.* **37**, e98499 (2018).
11. Henras, A. K., Plisson-Chastang, C., O'Donohue, M. F., Chakraborty, A. & Gleizes, P. E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip. Rev. RNA* **6**, 225–242 (2015).
12. Tafforeau, L. et al. The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of pre-rRNA processing factors. *Mol. Cell* **51**, 539–551 (2013).
13. Badertscher, L. et al. Genome-wide RNAi screening identifies protein modules required for 40S subunit synthesis in human cells. *Cell Reports* **13**, 2879–2891 (2015).
14. Larburu, N. et al. Structure of a human pre-40S particle points to a role for RACK1 in the final steps of 18S rRNA processing. *Nucleic Acids Res.* **44**, 8465–8478 (2016).
15. Oeffinger, M., Dlakic, M. & Tollervy, D. A pre-ribosome-associated HEAT-repeat protein is required for export of both ribosomal subunits. *Genes Dev.* **18**, 196–209 (2004).
16. McCaughan, U. M. et al. Pre-40S ribosome biogenesis factor Tsr1 is an inactive structural mimic of translational GTPases. *Nat. Commun.* **7**, 11789 (2016).
17. Sun, Q. et al. Molecular architecture of the 90S small subunit pre-ribosome. *eLife* **6**, (2017).
18. Strunk, B. S. et al. Ribosome assembly factors prevent premature translation initiation by 40S assembly intermediates. *Science* **333**, 1449–1453 (2011).
19. Ferreira-Cerca, S., Kiburu, I., Thomson, E., LaRonde, N. & Hurt, E. Dominant Rio1 kinase/ATPase catalytic mutant induces trapping of late pre-40S biogenesis factors in 80S-like ribosomes. *Nucleic Acids Res.* **42**, 8635–8647 (2014).
20. Woolls, H. A., Lamanna, A. C. & Karbstein, K. Roles of Dim2 in ribosome assembly. *J. Biol. Chem.* **286**, 2578–2586 (2011).
21. Turowski, T. W. et al. Rio1 mediates ATP-dependent final maturation of 40S ribosomal subunits. *Nucleic Acids Res.* **42**, 12189–12199 (2014).
22. Strunk, B. S., Novak, M. N., Young, C. L. & Karbstein, K. A translation-like cycle is a quality control checkpoint for maturing 40S ribosome subunits. *Cell* **150**, 111–121 (2012).

**Acknowledgements** The authors thank S. Rieder, H. Sieber and A. Gilmozzi for technical assistance, T. Fröhlich for mass-spectrometry analysis and E. Hurt, T. Becker and L. Kater for discussions and critical comments on the manuscript. This research was supported by grants from the Deutsche Forschungsgemeinschaft (SFB646, GRK1721 and FOR1805 to R.B.) and by an European Research Council (ERC) Advanced Grant (CRYOTRANSLATION) to R.B. M.A. is supported by a DFG fellowship through the Graduate School of Quantitative Biosciences Munich (QBM).

**Author contributions** M.A., J.C., O.B. and R.B. designed the study. M.A. generated stable cell lines and purified native complexes. J.C. cloned and purified biogenesis factors and conducted the reconstitution. M.A., J.C. and O.B. prepared the cryo-EM samples and O.B. collected cryo-EM data. M.A. and J.C. processed the cryo-EM data for their respective samples and J.C. built the molecular models with the help of M.A. M.A., J.C. and R.B. analysed the structures, interpreted results and all authors wrote the manuscript.

**Competing interests** The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0193-0>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0193-0>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to R.B.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

**Molecular cloning.** An N-terminal Strep-Flag tag was added to the multiple cloning site of the commercial vector pcDNA5/FRT/TO (Invitrogen). The region coding for PNO1 was amplified from human cDNA (Amsbio) using KOD DNA polymerase (Merck) and inserted into the modified pcDNA5/FRT/TO-StFLAG vector with BamHI and XhoI. DNA coding for human RIOK1, PNO1 and NOB1 was amplified from a HEK293T cell reverse transcription cDNA library and cloned into a modified pET 28a vector (SUMO tag).

**Generation of cell lines.** The generation of cell lines that stably express Strep-Flag-tagged PNO1 was adapted from previous work<sup>23</sup>. In brief, HEK Flp-In 293 T-Rex cells (Invitrogen) were grown to 70% confluency and transfected with pcDNA5/FRT/TO-StFLAG-PNO1 and pOG44 (Invitrogen) using Lipofectamine 3000 (Thermo Scientific). Selection was performed following the manufacturer's protocols with 150 µg ml<sup>-1</sup> hygromycin B (Thermo Scientific). Cells were maintained in DMEM (Thermo Scientific), containing 10% FBS, 100 µg ml<sup>-1</sup> hygromycin B, 10 µg ml<sup>-1</sup> blasticidin, and 1 × penicillin/streptomycin and GlutaMAX (Thermo Scientific). Cell lines were not authenticated or tested for mycoplasma contamination.

**Native complex preparation.** Native 40S precursors were purified from a stable HEK293 T-Rex Flp-In cell line as previously described<sup>6</sup>. Twenty-four hours before collection, PNO1 expression was induced with 1.6 µg ml<sup>-1</sup> tetracycline. Cells were collected in 0.025% trypsin/EDTA (Thermo Scientific), pelleted for 7 min at 1,800g and 4 °C and washed once with PBS. Cells were then lysed for 30 min on ice using lysis buffer (10 mM HEPES pH 7.6, 100 mM KCl, 2 mM MgCl<sub>2</sub>, 1 mM dithiothreitol (DTT), 0.5 mM NaF, 0.1 mM Na<sub>2</sub>V<sub>3</sub>O<sub>4</sub>, 0.5% NP-40 substitute, and 1 × protease inhibitor (Sigma Aldrich)). The lysate was cleared for 15 min at 4,400g and 4 °C before transfer to equilibrated StrepTactin XT affinity beads (IBA Lifesciences), followed by a 2 h incubation in an overhead rotator. Beads were collected in small columns and washed four times with buffer A (10 mM HEPES pH 7.6, 100 mM KCl, 2 mM MgCl<sub>2</sub>, 1 mM DTT, 0.5 mM NaF, 0.1 mM Na<sub>2</sub>V<sub>3</sub>O<sub>4</sub> and 1 × protease inhibitor). Bound material was then eluted five times with 30 min incubation in buffer A containing 50 mM D-biotin (Roth). Combined eluates were transferred to equilibrated anti-Flag affinity beads (Sigma-Aldrich), incubated for 2 h, followed by five washing steps with buffer A and elution of purified complexes by five times incubation for 30 min with 0.2 mg ml<sup>-1</sup> 3 × Flag peptide (Sigma Aldrich) in buffer A. Finally, the combined eluates were concentrated on 300 kDa molecular mass cut-off filters (Sartorius) and analysed for protein concentration on a NanoDrop photometer (Thermo Scientific). Nikkol was added to a final concentration of 0.05% to improve ice quality after grid preparation. Samples were then analysed by polyacrylamide gel electrophoresis on 4–12% Bis-Tris gradient gels (NuPAGE, Thermo Scientific), stained with SimplyBlue (Thermo Scientific) and bands were cut out and identified separately in-house via mass spectrometry.

**Reconstitution of pre-40S complexes.** For recombinant expression, NOB1, PNO1 and RIOK1 plasmids were transformed into *Escherichia coli* Bl21 strains and induced with 0.1 mM IPTG at 15 °C. Cells were lysed in 50 mM Tris pH 8.0, 300 mM NaCl, the lysate subjected to Ni-NTA affinity chromatography and proteins eluted overnight by on-column cleavage with Ulp1 at 4 °C. Samples were purified further using ion exchange and size exclusion chromatography with a final buffer condition of 20 mM Tris-HCl pH 8.0, 300 mM NaCl, and 5 mM DTT.

Human 40S subunits were purified from a HEK293T cell line. Cells were collected and disrupted using lysis buffer (20 mM HEPES pH 7.4, 100 mM potassium acetate, 7.5 mM magnesium acetate, 1 mM DTT, 10 µg cycloheximide, 0.5% NP-40 and 1 × protease inhibitor) and 80S ribosomes pelleted using a sucrose cushion (50 mM Tris pH 8.0, 500 mM potassium acetate, 25 mM magnesium acetate, 1 mM DTT, 1 M sucrose and 0.1% Nikkol). The resuspended ribosomes were stored in subunit separation buffer (50 mM HEPES pH 7.4, 500 mM KCl, 2 mM MgCl<sub>2</sub> and 2 mM DTT). To obtain separated 40S subunits, resuspended ribosomes were treated with puromycin at a final concentration of 1 mM, first 15 min on ice and then 10 min at 37 °C. After the treatment, the mixture was applied onto a 10–40% sucrose gradient and run overnight in a SW 40 rotor at 49,500g for 18 h. The 40S peak was collected and changed to storage buffer conditions (20 mM HEPES pH 7.4, 100 mM potassium acetate, 2.5 mM magnesium acetate and 2 mM DTT).

For the reconstitution of pre-40S complexes, purified small subunits (approximately 0.1 µM) were mixed with 2 µM RIOK1, 2 µM PNO1 and 2 µM NOB1 in storage buffer on ice for 30 min. Complexes were stabilized by crosslinking by adding glutaraldehyde to a final concentration of 0.5% (v/v). After 10 min incubation at room temperature, grids were prepared as described.

**Electron microscopy and image processing.** Pre-coated (2 nm) R3/3 holey carbon supported copper grids (Quantifoil) were glow discharged at 0.2 hPa for 20 s. Then, 3.5 µl of sample was directly applied onto each grid, blotted for 2–3 s at 4 °C and plunge frozen in liquid ethane using a Vitrobot Mark IV (FEI Company).

Grids were screened for ice quality and cryo-EM data acquired on a Titan Krios transmission electron microscope (FEI Company) at 300 kV under low-dose conditions (10 frames at about 2.5 e<sup>-</sup> Å<sup>-2</sup>) with a nominal pixel size of 1.084 Å per pixel on the object scale using the semi-automated software EM-TOOLS (TVIPS). A total of 17,345 and 11,977 micrographs of the native PNO1-pullout were collected on a Falcon II direct electron detector (datasets 1 and 2) at nominal defocus ranges from -1.0 to -2.5 µm. In total, 2,663 and 8,115 micrographs were collected of the two datasets 'mature 40S' (dataset 3) and 'reconstituted pre-40S' (dataset 4). Original image stacks were aligned, summed and drift-corrected using MotionCorr<sup>24</sup>. Contrast-transfer-function parameters and resolution were estimated for each micrograph using CTFFIND4<sup>25</sup> and Gctf<sup>26</sup>, respectively. Micrographs with an estimated resolution below 4 Å and astigmatism below 5% were manually screened for contamination or carbon rupture. Then, 11,356, 10,870, 2,523 and 7,753 micrographs were submitted to automated particle picking using Gautomatch (<https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/>) and human 40S (EMDB-5592) as reference, resulting in 1,154,906, 1,637,274, 407,657 and 959,348 picked particles. These were then reference-free 2D-classified and classes individually 3D refined as shown in Extended Data Fig. 1 using Relion (version 2.1-beta-1)<sup>27,28</sup>. Datasets 1 and 2 were individually processed before all particles of corresponding states were combined and further analysed (see Extended Data Fig. 1d).

**Model building and refinement.** In general, for the molecular model building of ribosome biogenesis factors, we initially performed secondary structure prediction (PSIPRED<sup>29</sup>) and, whenever crystal structures of homologues were available, we used 3D structure prediction by using SWISS-MODEL<sup>30</sup>. A full model was built for state C, while working models were prepared for all other states.

In class C, the *Saccharomyces cerevisiae* crystal structure (PDB code 5IW7) was used as a reference for TSR1, followed by manual refinement and de novo building of the missing extensions (47–75, 392–405 and 464–485) into the density in Coot<sup>31</sup>. For RIOK2, the *Chaetomium thermophilum* crystal structure (PDB code 4GYG) was used to generate a homology model, which was then rigid body fitted into its density and manually adjusted, before building de novo its C-terminal helices (494–544). The PNO1 structure is based on the *Pyrococcus horikoshii* crystal structure (PDB code 3AEV) with some minor adjustments to account for structural differences. For NOB1, the *P. horikoshii* NMR solution structure (PDB: code 2LCQ) was used as a reference to build a starting model for the PIN domain, followed by a manual refinement. The zinc-finger, insertion and C-terminal extension were all built de novo (94–233 and 256–412). The *S. cerevisiae* crystal structure (PDB code 5WVO) was used as a starting model for ENP1 and parts of LTV1. ENP1 and the N terminus of LTV1 were then manually adjusted and built into the map in Coot, respectively. The C-terminal helix of LTV1 (414–417) was built de novo. The 18S rRNA and associated ribosomal proteins were modelled using the human 40S ribosome (PDB code 5A2Q) as a reference, followed by rigid body fitting and manual adjustment. Immature h34 and h44, the linker region between h44 and h45, as well as the 3' end and ITS1 were built into their respective densities. The final model was real space refined at a resolution of 3.6 Å with secondary structure restraints for proteins and RNA, generated by ProSMART<sup>32</sup> and LIBG<sup>33</sup>, using PHENIX<sup>34</sup> and REFMAC5<sup>35</sup>. Final model evaluation was done with MolProbity<sup>36</sup>. Overfitting statistics were calculated by a random displacement of atoms in the model followed by a refinement with REFMAC5 against one of the half maps. Finally, Fourier shell correlation curves are calculated between the volume of the refined model and both half maps using Relion.

For state A, our model of state C was fitted into the density, before removing or adjusting parts to account for conformational and compositional differences. Namely, NOB1, LTV1, RACK1, RIOK2, as well as uS2, uS5, eS12, eS21, eS31 and parts of eS17 were removed from the model. A secondary structure model of RRP12 obtained in state B was placed in the respective density. Models for immature rRNA helices h34 and h39–40 were used from a model of state B and manually adjusted to their slightly different conformation. Helices h35–h37 from state C were rigid body fitted to additional densities enclosed by RRP12. A homology model of BUD23 and TRMT112 was generated based on PDB code 4QTU and rigid body placed into its respective density.

For state B, the model for state C was rigid body docked into the density. Well-resolved areas within the body were then Phenix real space refined, while less resolved areas within the head were solely manually checked. To account for structural differences, RACK1 and RIOK2, as well as eS12 and eS31 were removed from the structure. Flipped out 18S rRNA regions h34, h39 and h40 from state C were rigid body fitted together with the N-terminal part of eS17 into the focus-refined map in Chimera. RRP12 was modelled on a secondary structure level by manually placing poly-alanine alpha-helices in rod shaped densities.

For state D and E, rRNA and ribosomal proteins from the mature state, as well as NOB1, PNO1, RIOK2, LTV1 and TSR1 from state C were rigid body fitted into their respective density using Chimera.

A model for the mature 40S (PDB code 5A2Q) was fitted into and Phenix real space refined against the final mature 40S volume.



Finally, 18S rRNA and ribosomal proteins of the mature state were fitted into the reconstituted 40S precursor volume (state R). NOB1 and PNO1 were used from our model of state C and for RIOK1, only the C terminus which shows high resolution was de novo built, while the kinase domain was left out.

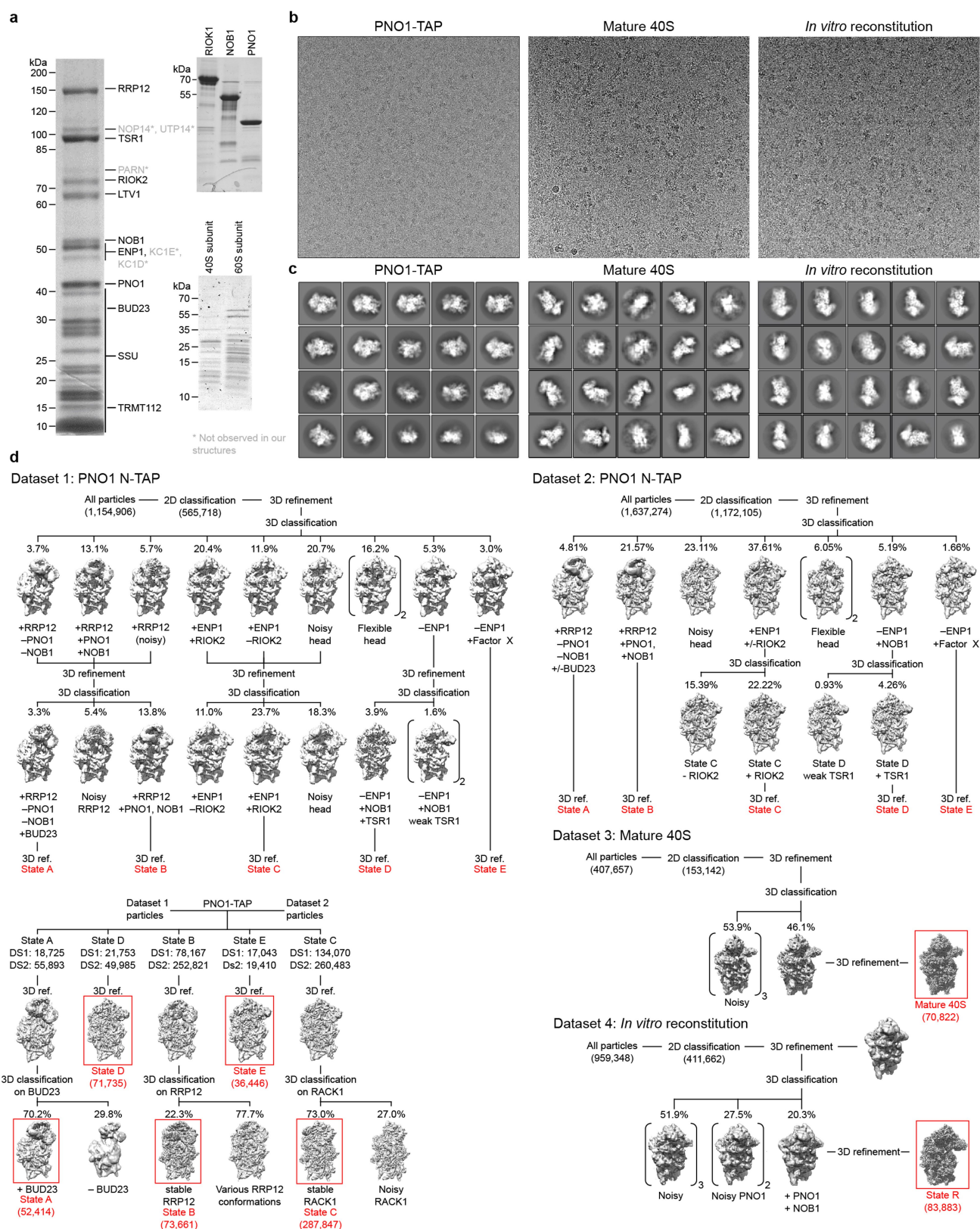
Maps and models were visualized and figures created with the PyMOL Molecular Graphics System (Version 1.7.4, Schrödinger, LLC), ChimeraX<sup>37</sup> and UCSF Chimera.

**Statistics and reproducibility.** Purification and sample preparation of native pre-40S complexes was done twice ( $n = 2$ ) with equal results. Cryo-EM data from two different grids has been collected with similar results (see Extended Data Fig. 1d). Reconstitution of 40S precursors and respective cryo-EM data collection has been done once ( $n = 1$ ). No statistical analysis has been applied throughout the work.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

**Data availability statement.** All cryo-EM density maps have been deposited in the Electron Microscopy Data Bank (EMDB) under the accession codes EMD-4349, EMD-4348, EMD-4337, EMD-4350, EMD-4351, EMD-4353 and EMD-4352. The atomic model of state C and all working models have been deposited in the Protein Data Bank (PDB) under accessions 6G4W, 6G4S, 6G18, 6G51, 6G53, 6G5I and 6G5H.

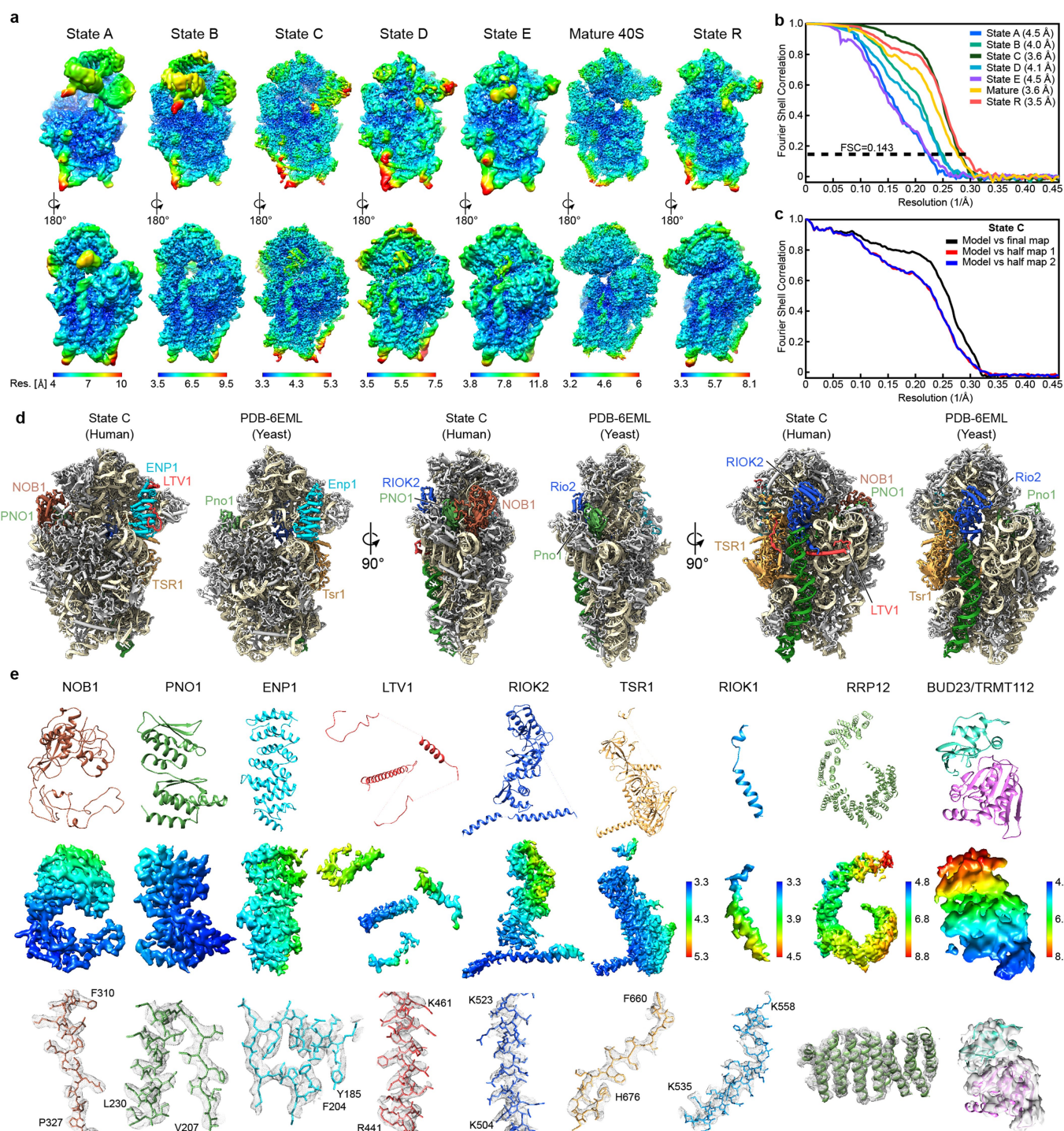
23. Glatzer, T., Wepf, A., Aebersold, R. & Gstaiger, M. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* **5**, 237 (2009).
24. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
25. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
26. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
27. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
28. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
29. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–W357 (2013).
30. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–W258 (2014).
31. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
32. Nicholls, R. A., Fischer, M., McNicholas, S. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr. D* **70**, 2487–2499 (2014).
33. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
34. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
35. Vagin, A. A. et al. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D* **60**, 2184–2195 (2004).
36. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
37. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).



**Extended Data Fig. 1 | Sample preparation and cryo-EM analysis of the pre-40S ribosome. a**, Coomassie stained SDS-PAGE analysis of samples used in this work with bands labelled as identified by mass spectrometry. RBFs that we do not observe in our structures are marked with an asterisk. 60S bands are shown for comparison (see Supplementary Fig. 1 for source data). **b**, Representative micrographs from three datasets, low-pass filtered at 15 Å. **c**, Representative 2D classes of three datasets showing various orientations of pre-40S particles. **d**, Cryo-EM data processing scheme for all datasets with final volumes highlighted in red. Classes that could

not be further refined or sorted are labelled flexible or noisy. For PNO1 N-TAP, two datasets were collected and individually processed. Particles of respective states were combined and further analysed (see Methods). Complex purification was done three times independently with the same results. Cryo-EM data collection and analysis was done twice for native complexes with similar results. Mass spectrometry analysis, pre-40S reconstitution and data collection for datasets 3 and 4 were done once. DS, dataset.

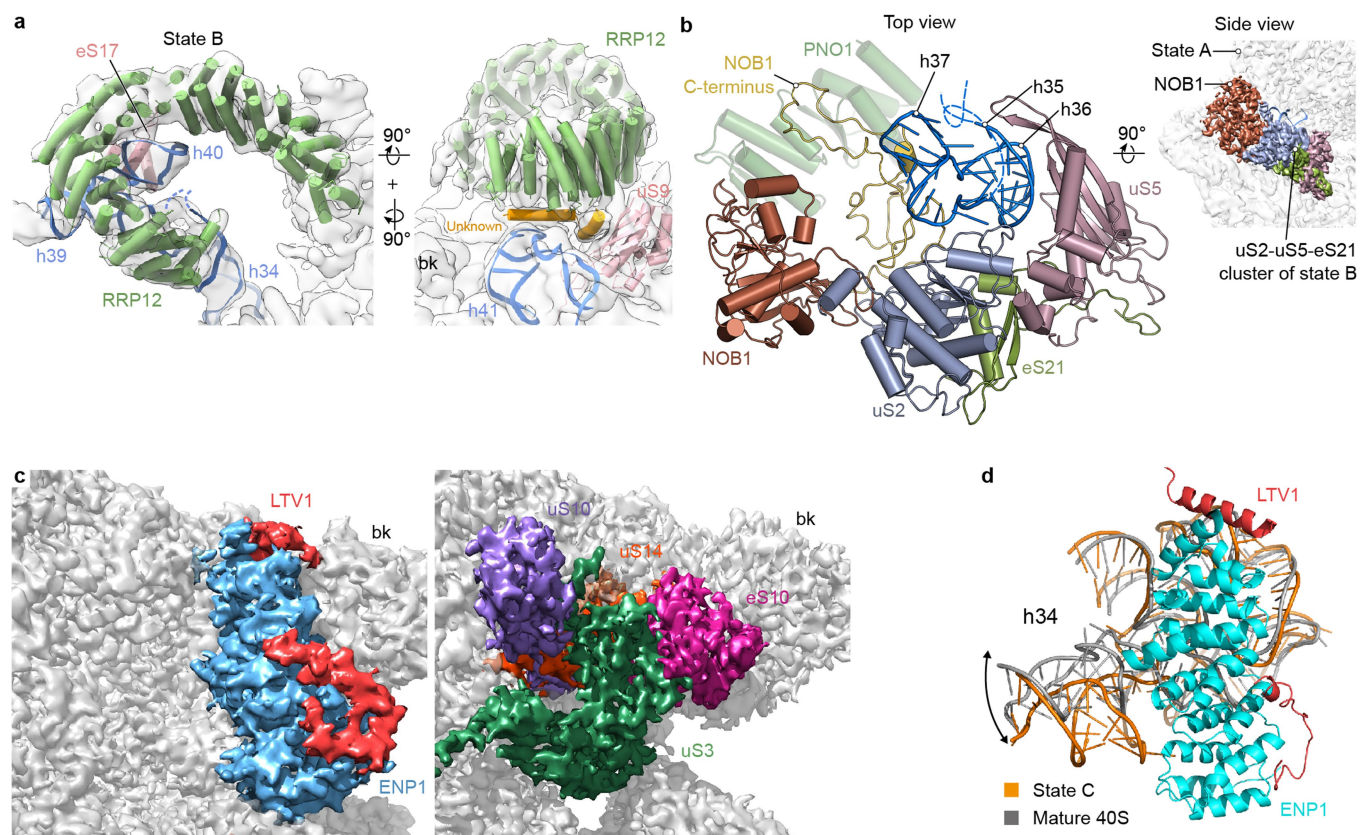




**Extended Data Fig. 2 | Local resolution, refinement and model statistics.** **a**, Local resolution distribution as estimated by Relion ranging from approximately 3 Å in well-resolved areas to 12 Å in more flexible parts. Colouring according to scale bars. **b**, Fourier shell correlation (FSC) plot with average resolutions according to the 'gold standard' (FSC = 0.143) stated in the legend. **c**, FSC plot of the state C model against cryo-EM maps as calculated by REFMAC5 (see Supplementary Data 1 for source data). **d**, Structural comparison between state C and a pre-40S

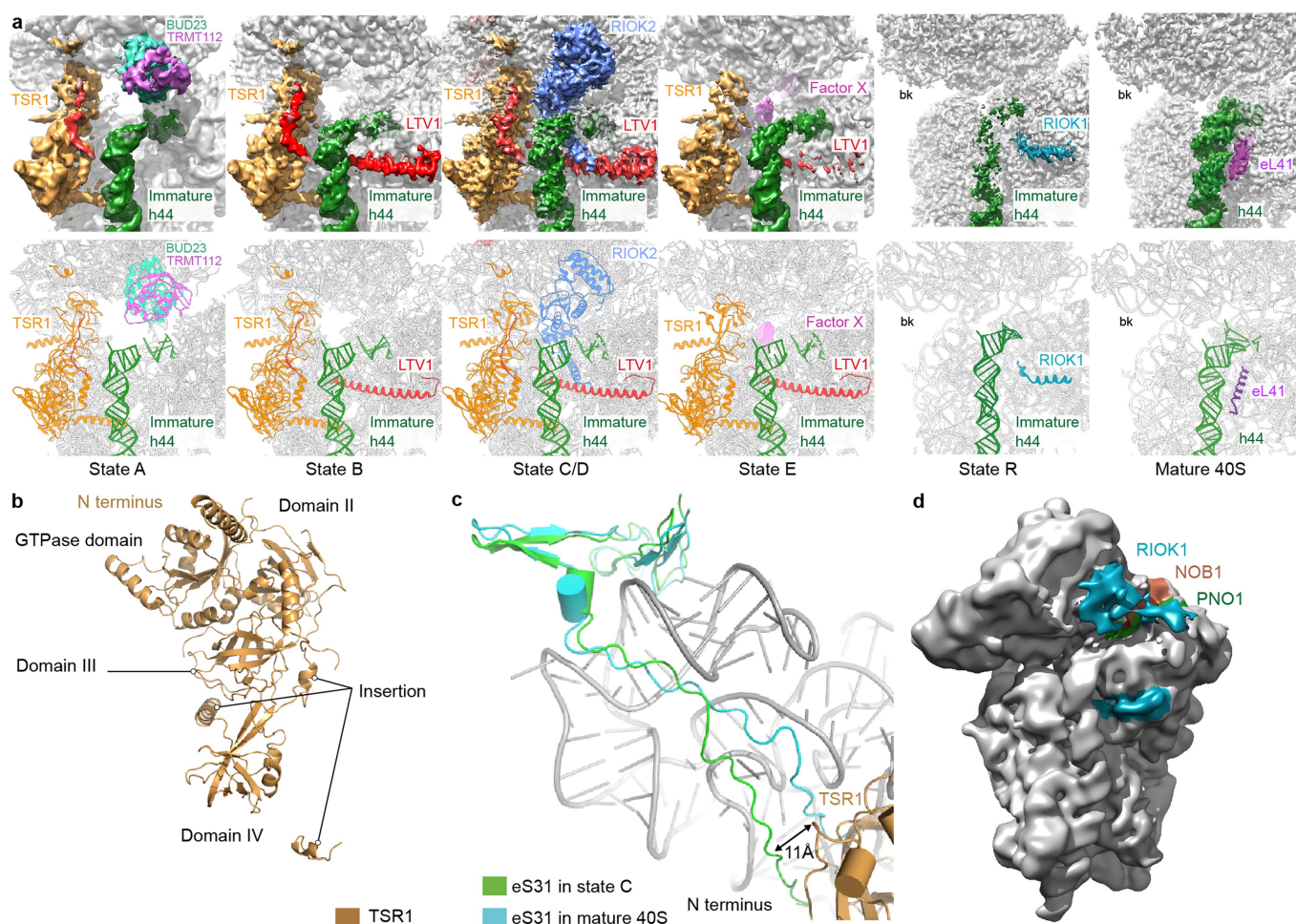
particle from yeast (PDB code 6EML). Assembly factors coloured as in Fig. 1a. **e**, Cartoon representations of models of RBFs (top) are shown together with their respective density (mid). Volumes are coloured according to local resolution, which ranges from 3 Å in more rigid areas to 9 Å in flexible parts. RRP12, BUD23 and TRMT112 are less resolved, which only permitted placing of dummy helices and rigid body fitting of homology models, respectively. Examples of well resolved areas are depicted below.





**Extended Data Fig. 3 | Structural details of assembly factors and beak formation.** **a**, Positioning of RRP12 in state B shows interaction with eS17 and uS9, as well as rRNA h34, h39, h40 and h41. Two additional helices of unknown identity bridge RRP12 and the head. Poly-alanine helices are placed in respective densities. **b**, Recruitment of NOB1 and the

uS2-u5-eS21 cluster after h35-h37 flipping; eS17 is omitted from this view. **c**, Close-up view of the beak region before (left) and after (right) replacement of ENP1-LTV1 with the uS3-eS10-uS20 cluster. **d**, Movement of h34 during maturation.

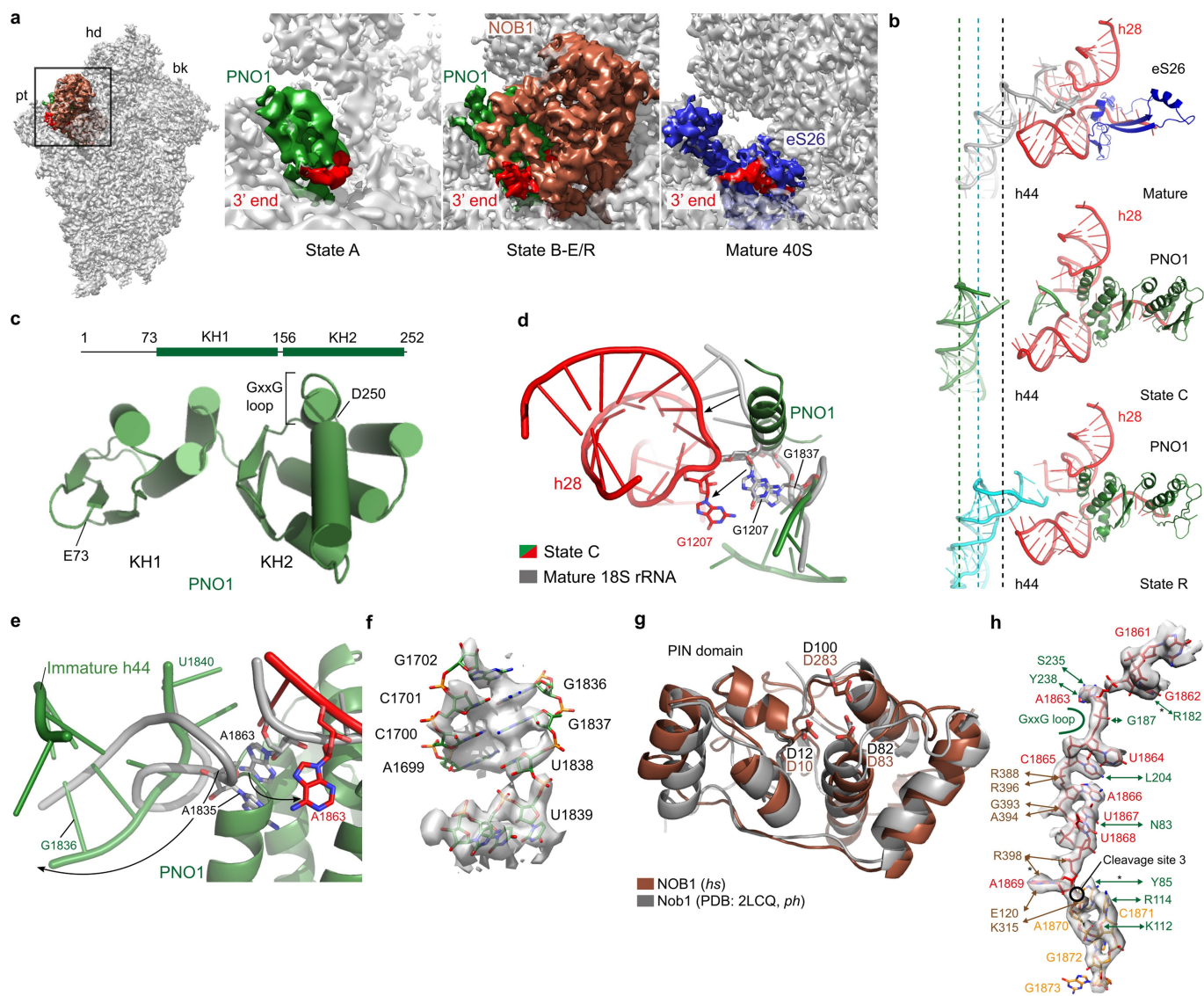


#### Extended Data Fig. 4 | The maturation of the decoding centre.

**a**, Close-up views of the decoding centre in density (top) and model (bottom) representation showing the position of the biogenesis factors close to h44. **b**, Domain arrangement of TSR1. Insertions and extensions

of TSR1 are labelled. **c**, Conformational change of the N terminus of eS31 during maturation. The alignment was based on the C-terminal zinc-finger domain. **d**, Overall structure of the reconstituted particle low-pass filtered at 12 Å, showing the approximate positioning of ROK1 (blue).

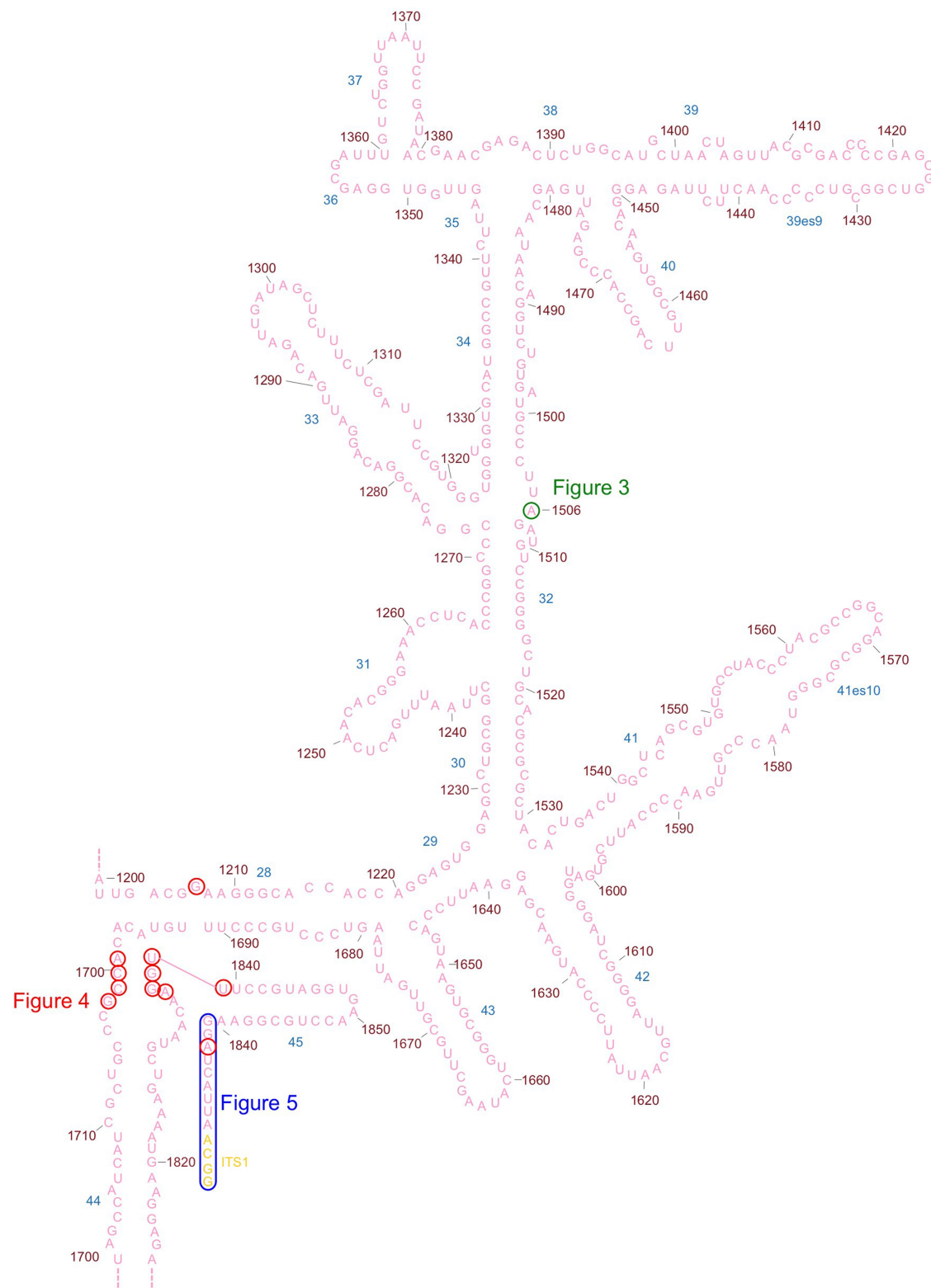




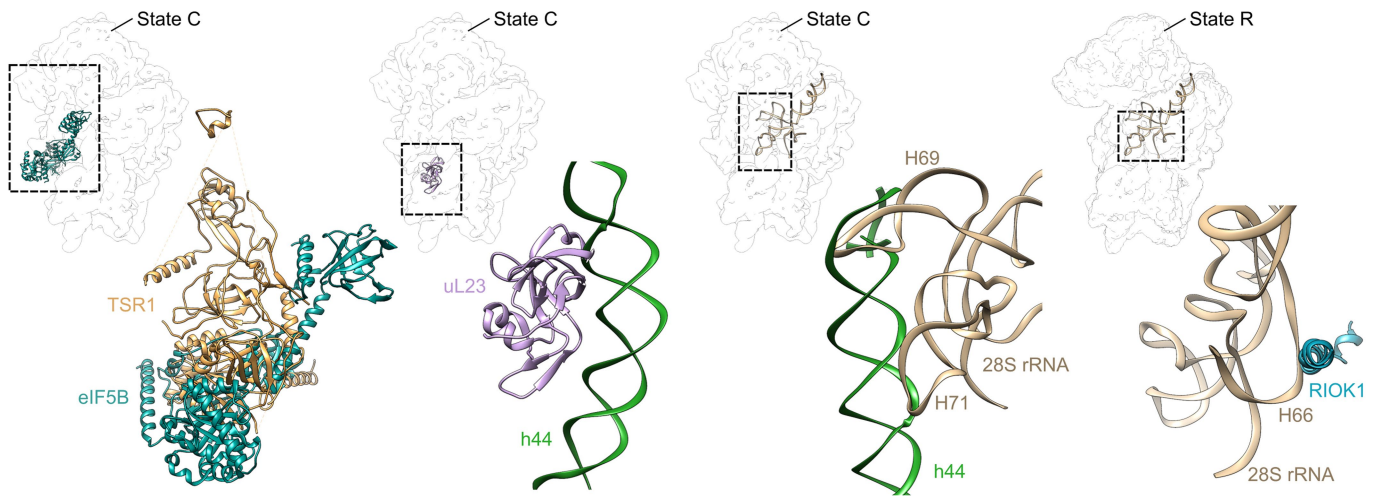
**Extended Data Fig. 5 | PNO1, NOB1 and coordination of the 3' rRNA end.** **a**, The 3' end region (box) throughout the maturation process. PNO1 and NOB1 bind similarly in state B–E and R. **b**, Cartoon representation of eS26 and PNO1 with parts of the pre-18S rRNA at different stages in maturation. h44 is shifted in state C. Displacement of eS26 by PNO1 in state R leads to a partially shifted h44. **c**, Overall structure of PNO1 with its two KH domains. **d**, **e**, Detailed view on the effects of PNO1 binding. Clashing of PNO1 with h28 and the linker region between h44 and h45

leads to a disruption of the base stacking between G1207 and G1837 (**d**) and A1835 and A1863 (**e**). **f**, Residues involved in novel base pairing with their respective electron density in state C. **g**, Comparison of the PIN domain of human (*hs*) NOB1 in state C and the NMR structure of Nob1 from *Pyrococcus horikoshii* (*ph*; PDB code 2LCQ), with their conserved active site residues highlighted. **h**, Summary of residues involved in binding of the rRNA 3' end (asterisk indicates stacking). Electron density of state C surrounding the 3' end is shown.





**Extended Data Fig. 6 | 2D diagram of the pre-18S rRNA head region of state C.** Extended secondary structure diagram of the pre-18S rRNA head region of state C. Residues mentioned throughout the text are highlighted. Related to Fig. 4b.



**Extended Data Fig. 7 | 60S subunits components clash with immature h44.** Overview over clash sites of large subunit components with pre-40S particles. Left, initiation factor eIF5B, which is potentially involved in formation of the 80S-like ribosome complex, occupies a similar binding site as TSR1. Middle, ribosomal protein uL23 and helices H69 and H71 of

28S rRNA clash with h44 in its immature position. Right, finally helix H66 of 28S rRNA clashes with a R1OK1 helix. A model of an 80S ribosome with eIF5B (PDB code 4UJD) was aligned to pre-40S states, and the factors are shown as overlays together with one of the 40S precursors.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	<b>PNO1-TAP (EMD-4437) (PDB-6G18)</b>	<b>State R (EMD-4353) (PDB-6G5I)</b>	<b>Mature 40S (EMD-4352) (PDB-6G5H)</b>
<b>DATA COLLECTION</b>			
Magnification	129,151	129,151	129,151
Voltage (kV)	300	300	300
Electron exposure (e <sup>-</sup> Å <sup>-2</sup> )	25	25	25
Defocus range (μm)	-0.4 to -3.2	-0.7 to -3.3	-0.7 to -3.8
Pixel size (Å)	1.084	1.084	1.084
Symmetry imposed	C1	C1	C1
Initial particle images (no.)	1,737,823	959,348	407,657
Final particle images (no.)	287,847 (state C)	83,883	70,822
Map resolution (Å)	3.6 (state C)	3.5	3.6
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)	3.3-7.5	3.3-6.8	3.4-8.6
<b>MODEL REFINEMENT</b>			
<b>Refinement</b>			
Initial model used	5A2Q		
Model resolution (Å)	3.6		
FSC threshold	0.5		
Map sharpening <i>B</i> factor (Å <sup>2</sup> )	-210.9		
Model resolution range (Å)	3.3-7.5		
<b>Model composition</b>			
Non-hydrogen atoms	83825		
Protein residues	6060		
RNA bases	1658		
Ligands	2		
<i>B</i> factors (Å <sup>2</sup> )	141.1		
<b>R.m.s. deviations</b>			
Bond lengths (Å)	0.0068		
Bond angles (°)	1.19		
<b>Validation</b>			
Molprobity score	2.46		
Clashscore, all atoms	6.28		
Poor rotamers (%)	7.77		
<b>Ramachandran plot</b>			
Favoured (%)	93.33		
Allowed (%)	5.85		
Outliers (%)	0.82		
<b>Validation (RNA)</b>			
Correct sugar puckers (%)	96.14		
Good backbone conformations (%)	67.37		

Summary of relevant parameters used during cryo-EM data collection and processing. Refinement and validation statistics are provided for the molecular model of state C.



# Stellar populations dominated by massive stars in dusty starburst galaxies across cosmic time

Zhi-Yu Zhang<sup>1,2</sup>, D. Romano<sup>3</sup>, R. J. Ivison<sup>1,2\*</sup>, Padelis P. Papadopoulos<sup>4,5</sup> & F. Matteucci<sup>6,7,8</sup>

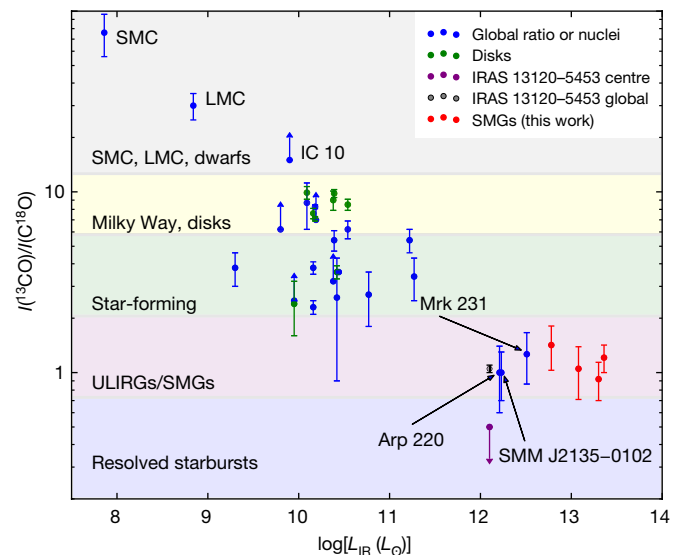
All measurements of cosmic star formation must assume an initial distribution of stellar masses—the stellar initial mass function—in order to extrapolate from the star-formation rate measured for typically rare, massive stars (of more than eight solar masses) to the total star-formation rate across the full stellar mass spectrum<sup>1</sup>. The shape of the stellar initial mass function in various galaxy populations underpins our understanding of the formation and evolution of galaxies across cosmic time<sup>2</sup>. Classical determinations of the stellar initial mass function in local galaxies are traditionally made at ultraviolet, optical and near-infrared wavelengths, which cannot be probed in dust-obscured galaxies<sup>2,3</sup>, especially distant starbursts, whose apparent star-formation rates are hundreds to thousands of times higher than in the Milky Way, selected at submillimetre (rest-frame far-infrared) wavelengths<sup>4,5</sup>. The  $^{13}\text{C}/^{18}\text{O}$  isotope abundance ratio in the cold molecular gas—which can be probed via the rotational transitions of the  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  isotopologues—is a very sensitive index of the stellar initial mass function, with its determination immune to the pernicious effects of dust. Here we report observations of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  emission for a sample of four dust-enshrouded starbursts at redshifts of approximately two to three, and find unambiguous evidence for a top-heavy stellar initial mass function in all of them. A low  $^{13}\text{CO}/\text{C}^{18}\text{O}$  ratio for all our targets—alongside a well tested, detailed chemical evolution model benchmarked on the Milky Way<sup>6</sup>—implies that there are considerably more massive stars in starburst events than in ordinary star-forming spiral galaxies. This can bring these extraordinary starbursts closer to the ‘main sequence’ of star-forming galaxies<sup>7</sup>, although such main-sequence galaxies may not be immune to changes in initial stellar mass function, depending on their star-formation densities.

Oxygen, carbon and their stable isotopes are produced solely by nucleosynthesis in stars<sup>8</sup>. The minor isotopes,  $^{13}\text{C}$  and  $^{18}\text{O}$ , are released mainly by low- and intermediate-mass stars (those with stellar mass less than eight solar masses,  $M_* < 8M_\odot$ ) and massive stars ( $M_* > 8M_\odot$ ), respectively<sup>9</sup>, owing to their differing energy barriers in nuclear reactions and evolution of stars<sup>10</sup>. These isotopes then mix with the interstellar medium (ISM) such that the  $^{13}\text{C}/^{18}\text{O}$  abundance ratio measured in the ISM becomes a ‘fossil’, imprinted by evolutionary history and the stellar initial mass function (IMF)<sup>6</sup>. The abundances of the  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  isotopologues in the molecular ISM, whose measurements are immune to the pernicious effects of dust, are therefore a very sensitive index of the IMF in galaxies.

Galaxies in the early Universe, having had much less cosmological time available for prior episodes of evolution, are expected to have simpler star-formation histories than local galaxies. Our sample comprises the strongest carbon monoxide (CO) emitters in the early Universe, selected from the literature (see Methods): four gravitational lensed submillimetre galaxies at redshift  $z \approx 2\text{--}3$ , with look-back times exceeding ten billion years.

Using the Atacama Large Millimeter Array (ALMA) telescope, we have robustly ( $>5\sigma$ , where  $\sigma$  is the standard deviation) detected multiple transitions of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  in most of our target galaxies. The  $J=3 \rightarrow 2$  lines from galaxy SDP.17b and the  $J=5 \rightarrow 4$  lines from galaxy SPT 0103–45 are marginally detected at approximately  $4\sigma$  levels. But the  $J=4 \rightarrow 3$  transitions of SDP.17b are detected at high signal-to-noise ratio, so we can be confident that emission features seen at the expected velocities of the weaker transitions are also real. We also detected  $^{12}\text{CO } J=4 \rightarrow 3$  and  $J=5 \rightarrow 4$  for SPT 0125–47 and SPT 0103–45, respectively.

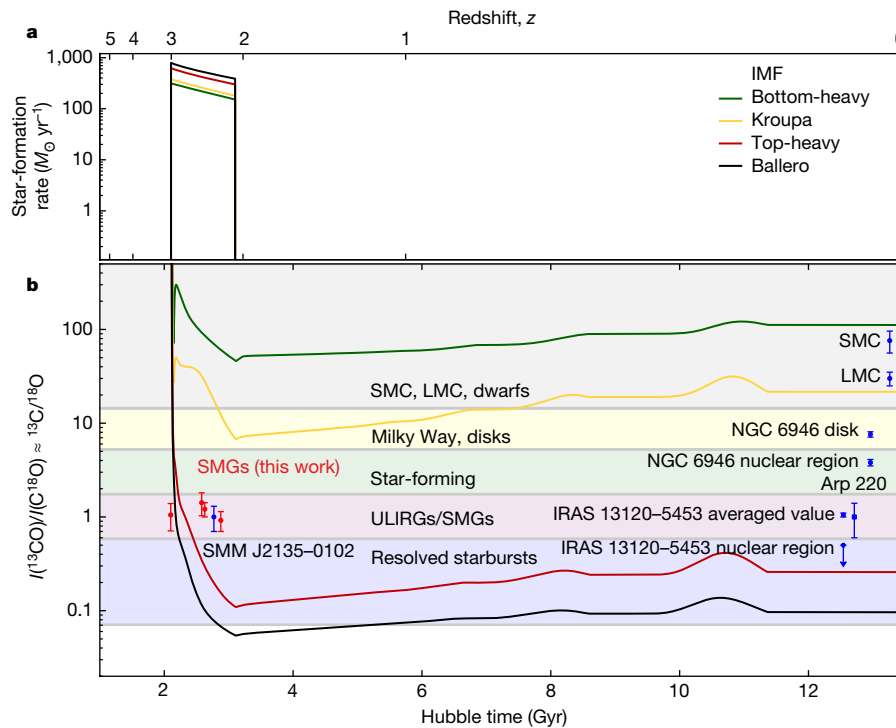
As shown in Fig. 1, there is a decreasing trend in the ratio of velocity-integrated line intensities,  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$ , with increasing infrared luminosity,  $L_{\text{IR}}$  (or the apparent star-formation rate traced by massive stars). For all the galaxies in our observed sample (see Methods), the line ratios of  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  are close to unity, similar to those found<sup>11,12</sup> in three nearby ultraluminous infrared galaxies (ULIRGs,



**Fig. 1 |  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  as a function of  $L_{\text{IR}}$ , corrected for gravitational amplification when appropriate.** The rest-frame for  $L_{\text{IR}}$  is  $8\text{--}1,000\ \mu\text{m}$ . Red symbols refer to submillimetre galaxies (SMGs) in our sample. We include  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  measurements of the Milky Way, nearby star-forming spiral galaxies, disks, a few resolved nuclei<sup>15</sup>, three local ULIRGs<sup>11</sup>, Arp 220, Mrk 231, and IRAS 13120–5453, and a submillimetre galaxy<sup>13</sup>, SMM J2135–0102 at  $z \approx 2.3$ . The ratios of the Small and Large Magellanic Clouds (SMC and LMC) are averaged from multiple positions<sup>58,29</sup>.  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  in dwarf galaxy IC 10 is reported as a lower limit<sup>30</sup>. A decreasing trend of  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  as a function of  $L_{\text{IR}}$  is clearly evident, indicating that  $^{13}\text{C}/^{18}\text{O}$  abundance ratios are varying systematically in galaxies with different rates of apparent star formation. Error bars represent  $1\sigma$  uncertainty.

<sup>1</sup>Institute for Astronomy, University of Edinburgh, Edinburgh, UK. <sup>2</sup>European Southern Observatory, Garching, Germany. <sup>3</sup>INAF, Astrophysics and Space Science Observatory, Bologna, Italy.

<sup>4</sup>Department of Physics, Section of Astrophysics, Astronomy and Mechanics, Aristotle University of Thessaloniki, Thessaloniki, Greece. <sup>5</sup>Research Center for Astronomy, Academy of Athens, Athens, Greece. <sup>6</sup>Department of Physics, Section of Astronomy, University of Trieste, Trieste, Italy. <sup>7</sup>INAF, Osservatorio Astronomico di Trieste, Trieste, Italy. <sup>8</sup>INFN, Sezione di Trieste, Trieste, Italy. <sup>9</sup>e-mail: rob.ivison@eso.org



**Fig. 2 | Theoretical  $^{13}\text{C}$  and  $^{18}\text{O}$  isotopic abundance ratios in the ISM for different evolutionary tracks, predicted using various IMFs.**

**a**, Star-formation history for a delayed starburst, starting two billion years after the Big Bang, with a total cessation of subsequent star formation. Coloured lines correspond to different IMFs. **b**, Theoretical

$^{13}\text{C}/^{18}\text{O}$  abundance ratio in the ISM as a function of time, following the different IMFs (shown in colour). Red symbols refer to submillimetre galaxies (SMGs) in our sample. Blue symbols show the  $^{13}\text{CO}/\text{C}^{18}\text{O}$  ratios measured in SMM J2135–0102 and a few representative local galaxies. Table 1 lists the detailed definitions of the IMFs adopted here.

with  $L_{\text{IR}} \geq 10^{12} L_{\odot}$ )—Arp 220, Mrk 231 and IRAS 13120–5453—as well as in the strongly lensed submillimetre galaxy<sup>13</sup> SMM J2135–0102 at redshift  $z \approx 2.3$ . Galactic disks of nearby spiral galaxies have  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  ratios similar to the representative ratio<sup>8,14</sup> of the Milky Way’s disk, about 7–10. In the central nuclear regions of these spiral galaxies, where the star-formation activity is more intense than in the disks,  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  ratios are lower<sup>15</sup>, though they remain restricted to  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O}) \geq 4$ . The Magellanic Clouds—our nearest dwarf galaxies—show the highest  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  ratios of about 30–60.

For representative Galactic abundance ratios of  $^{13}\text{CO}/\text{C}^{18}\text{O}$  of about 7–10, the  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  line intensity ratio can approach values near unity—which is what we measure for all the galaxies in our sample—only if even the rarest of our three isotopologue lines,  $\text{C}^{18}\text{O}$ , were to acquire substantial optical depths on galactic scales (see Methods). On the other hand, to reach line ratios  $I(^{12}\text{CO})/I(^{13}\text{CO})$  and  $I(^{12}\text{CO})/I(\text{C}^{18}\text{O})$  in excess of 30—as found in our sample—the optical depths of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  have to be much less than one for either type of condition—local thermodynamic equilibrium (LTE) or non-LTE excitation (see Methods)—assuming the typical abundance ratios of  $^{12}\text{CO}/^{13}\text{CO} \approx 40$ –100 found in the Milky Way<sup>6,8</sup>.

The magnification factors of gravitational lensing in our objects are modest ( $\mu \approx 5$ , see Extended Data Table 1), with the notable exception of the Cloverleaf quasar ( $\mu \approx 10$ ). It is unlikely that differential lensing could skew the measured  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  line ratio away from the value intrinsic to the galaxy, even in cases of much stronger lensing<sup>13</sup>. For differential lensing to operate in this way, the global  $I(^{13}\text{CO})$  and  $I(\text{C}^{18}\text{O})$  distributions over the galaxies must be very different, which is improbable given that these two isotopologue lines have almost identical excitation requirements and any differences in their distribution are expected to be confined within individual molecular clouds (see Methods). Finally, the isotopologue lines have been observed simultaneously, making the uncertainties from pointing and calibration negligible. Furthermore, it was recently shown that known

photo-chemical effects, such as selective photodissociation and fractionation, cannot induce global isotopologue abundances to differ from the intrinsic, IMF-determined, isotopic abundances in star-forming galaxies<sup>6,12</sup>.

We thus conclude that emission in both  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  is optically thin for the bulk of the molecular gas mass in these galaxies. The systematically low  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  ratios found in all our high-redshift starbursts—as well as in local ULIRGs—reflect intrinsic isotopologue abundance ratios over galaxy-sized molecular hydrogen reservoirs, that is,  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O}) \approx ^{13}\text{CO}/\text{C}^{18}\text{O} \approx ^{13}\text{C}/^{18}\text{O}$ . Figure 1 thus reflects a strong decrease of the  $^{13}\text{C}/^{18}\text{O}$  abundance ratio in starburst galaxies, compared to local spiral galaxies, galactic disks and dwarf galaxies.

The only plausible basis for such systematic variations of isotopologue abundance ratios over galaxy-sized molecular hydrogen reservoirs is a change of the stellar IMF, which must cause the intrinsic abundances of isotopic elements to deviate substantially from those found in ordinary star-forming systems where the standard IMF prevails. The robustness of this conclusion is made possible by tremendous advances in chemical evolution modelling, which now takes into account up-to-date stellar isotopic yields in dependence on both stellar masses and metallicities, time differentials for their release into the ISM, and the dependence of stars’ initial chemical composition on prior galactic evolution. Benchmarked against the rich isotopic datasets of the Milky Way, these models can now follow the chemical evolution of various isotopes and their abundance ratios, uniquely identifying the effects of different IMFs upon them<sup>6</sup> (see Methods).

In Fig. 2, we present chemical evolution models that show how the isotopologue abundance ratios are altered by IMF types, and how they evolve as a function of cosmic time. The models show a massive galaxy that started an intense burst of star formation at  $z \approx 3$ , reached a stellar mass of  $10^{11} M_{\odot}$  one billion years later, ceased forming stars after the burst, then evolved passively to the present day. This represents an extreme case for the evolution of the  $^{13}\text{C}/^{18}\text{O}$  abundance ratio for a pure starburst. The  $^{13}\text{C}/^{18}\text{O}$  abundance ratio starts from a high value

**Table 1 | Characteristics of the IMFs used in this work**

IMF	$\alpha_0$	$\alpha_1$	$\alpha_2$	$m_0$ ( $M_\odot$ )	$m_1$ ( $M_\odot$ )	$m_2$ ( $M_\odot$ )	$M_3$ ( $M_\odot$ )	$M_*^{(8-100)M_\odot} / M_*^{\text{total}}$ (%)
Bottom-heavy	-1.7	-1.7	-1.7	0.1	0.5	1.0	100	3.9
Kroupa <sup>3,6</sup>	-0.3	-1.2	-1.7	0.1	0.5	1.0	100	6.9
Top-heavy	-0.3	-1.1	-1.1	0.1	0.5	1.0	100	33.3
Ballero <sup>16</sup>	-0.3	-0.95	-0.95	0.1	0.5	1.0	100	44.0

The slopes quoted in the table are for IMFs in mass, where  $m_0$  ( $=0.1M_\odot$ ) and  $m_3$  ( $=100M_\odot$ ) are, respectively, the lower and upper limits of stellar masses assumed in the models, that is, the IMF is normalized to unity in the  $(0.1-100)M_\odot$  range.  $m_1$  and  $m_2$  indicate the masses at which there is a change in the IMF slope, if any. For instance, for the Kroupa IMF, the slope changes at both  $m_1$  and  $m_2$ ; for the Ballero and top-heavy IMFs, the slope changes only at  $m_1$ ; finally, the bottom-heavy IMF has a single slope. The Kroupa IMF slopes are adopted for reproducing typical Milky Way values in chemical evolution models<sup>6</sup>, which are within the error bars of the original reported values<sup>3</sup>.

set by the first generations of metal-poor massive stars, because  $^{13}\text{C}$  is released from its primary and secondary nucleosynthesis channels (see Methods) earlier than  $^{18}\text{O}$ , which is purely a secondary element (see Methods). The ratio drops quickly during the starburst, then slowly increases with time, varying by a factor of 2–3 depending on the adopted IMF and the time interval. The late increase of the ratio is due mostly to the slow but continuous release of  $^{13}\text{C}$  from low- and intermediate-mass stars (see Methods), which keep releasing  $^{13}\text{C}$  for a long time after star formation—and, at the same time, the  $^{18}\text{O}$  pollution from massive stars—has ceased; wiggles in the  $^{13}\text{C}/^{18}\text{O}$  abundance ratio correspond to the lifetimes of roughly solar-mass stars.

It is not possible to reproduce the observed  $I(^{13}\text{CO})/I(^{18}\text{O})$  ratio in submillimetre galaxies with a Kroupa IMF (or similar IMFs; see Table 1), that is, with an IMF that can reproduce the ratios found in the Milky Way and in the disks of local spiral galaxies<sup>3</sup>. The top-heavy IMF and the Ballero<sup>16</sup> IMF (which can reproduce the chemical abundances of stars in the Galactic bulge) under-produce the  $I(^{13}\text{CO})/I(^{18}\text{O})$  ratios observed in  $z \approx 2-3$  starburst galaxies and local ULIRGs, but they can reproduce the extremely low  $I(^{13}\text{CO})/I(^{18}\text{O})$  ratio recently measured<sup>12</sup> in the central 500-parsec region of the starburst ULIRG, IRAS 13120–5453, which has been until now the lowest value reported in the literature. Note that the average star-formation event may have a less top-heavy IMF over galactic scales, or a mix of both top-heavy and canonical IMFs that produces a galaxy-sized average  $^{13}\text{C}/^{18}\text{O} \leq 1$ , which also applies for the resolved studies of IRAS 13120–5453.

A clear trend is shown in Fig. 2: the more top-heavy the IMF, the lower the  $I(^{13}\text{CO})/I(^{18}\text{O})$  ratio, which is also compatible with the ratios found in local ULIRGs and the exceptionally small  $I(^{13}\text{CO})/I(^{18}\text{O})$  ratio found<sup>12</sup> in the centre of IRAS 13120–5453. This paints a consistent picture in which a top-heavy IMF operates within both local ULIRGs and the much more numerous, distant starburst galaxies, where starburst events can quickly enrich the  $^{18}\text{O}$  abundance, pushing the  $^{13}\text{C}/^{18}\text{O}$  ratio to (or below) unity. A canonical IMF can never produce a  $I(^{13}\text{CO})/I(^{18}\text{O})$  ratio close to unity, no matter what type of star-formation history or at what time along a galaxy's evolutionary track the measurement is made.

Multiple evidence in the local Universe has shown that the stellar IMF in galaxies with very high star-formation rate densities seems to be biased towards massive stars, such as ultra-compact dwarf galaxies<sup>17</sup>, ULIRGs<sup>18</sup> and progenitors of early-type galaxies<sup>19</sup>. A top-heavy stellar IMF was recently also found in compact stellar associations in the Large Magellanic Cloud<sup>20–22</sup>, whose high-density star-formation events may closely replicate what happens over galactic scales in distant starbursts. Our results—for the most intensive star-forming systems in the distant Universe, where classical ultraviolet and optical methods cannot be applied—are in line with these findings. We also note that metal-poor dwarf galaxies are likely to have an IMF biased towards low-mass stars, which is predicted by the integrated galaxy-wide IMF theory and is consistent with the results found in dwarf galaxies<sup>23</sup> and the outer regions of disk galaxies, using H $\alpha$  and ultraviolet observations<sup>24</sup>.

An IMF biased towards massive stars implies that the star-formation rates determined for submillimetre galaxies must be considerably reduced, since they are based on extrapolations of observables related to massive stars<sup>1</sup>. Moving from the Kroupa IMF to the Ballero IMF, the relative mass fraction of massive stars increases by a factor of 6–7

(see Table 1), meaning that star-formation rates derived from most classical tracers<sup>1</sup> (the  $L_{\text{IR}}$ , the radio continuum and so on) must decrease by a similar factor. As a result, dusty starburst galaxies probably lie much closer to the so-called ‘main sequence’ of star-forming galaxies<sup>7</sup> than previously thought. Classical ideas about the evolutionary tracks of galaxies<sup>25</sup> and our understanding of cosmic star-formation history<sup>26</sup> are challenged. Fundamental parameters governing galaxy formation and evolution—star-formation rates, stellar masses, gas-depletion and dust-formation timescales, dust extinction laws, and more<sup>27</sup>—must be re-addressed, exploiting recent advances in stellar physics.

### Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0196-x>.

Received: 29 September 2017; Accepted: 27 February 2018;

Published online: 04 June 2018

- Kennicutt, R. C. Jr. Star formation in galaxies along the Hubble sequence. *Annu. Rev. Astron. Astrophys.* **36**, 189–231 (1998).
- Bastian, N., Covey, K. R. & Meyer, M. R. A universal stellar initial mass function? A critical look at variations. *Annu. Rev. Astron. Astrophys.* **48**, 339–389 (2010).
- Kroupa, P. et al. *The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations* Ch. 4, 115–242 (Springer, Dordrecht, 2013).
- Smail, I., Ivison, R. J. & Blain, A. W. A deep sub-millimeter survey of lensing clusters: a new window on galaxy formation and evolution. *Astrophys. J.* **490**, L5–L8 (1997).
- Hughes, D. H. et al. High-redshift star formation in the Hubble Deep Field revealed by a submillimetre-wavelength survey. *Nature* **394**, 241–247 (1998).
- Romano, D., Matteucci, F., Zhang, Z.-Y., Papadopoulos, P. P. & Ivison, R. J. The evolution of CNO isotopes: a new window on cosmic star formation history and the stellar IMF in the age of ALMA. *Mon. Not. R. Astron. Soc.* **470**, 401–415 (2017).
- Noeske, K. G. et al. Star formation in AEGIS field galaxies since  $z=1.1$ : the dominance of gradually declining star formation, and the main sequence of star-forming galaxies. *Astrophys. J.* **660**, L43–L46 (2007).
- Wilson, T. L. & Matteucci, F. Abundances in the interstellar medium. *Astron. Astrophys. Rev.* **4**, 1–33 (1992).
- Romano, D., Karakas, A. I., Tosi, M. & Matteucci, F. Quantifying the uncertainties of chemical evolution studies. II. Stellar yields. *Astron. Astrophys.* **522**, A32 (2010).
- Pagel, B. E. J. *Nucleosynthesis and Chemical Evolution of Galaxies* (Cambridge Univ. Press, Cambridge, 2009).
- Henkel, C. et al. Carbon and oxygen isotope ratios in starburst galaxies: new data from NGC 253 and Mrk 231 and their implications. *Astron. Astrophys.* **565**, A3 (2014).
- Slawi, K., Wilson, C. D., Aalto, S., Privon, G. C. & Extreme, C. O. Isotopic abundances in the ULIRG IRAS 13120–5453: an extremely young starburst or top-heavy initial mass function. *Astrophys. J.* **840**, L11 (2017).
- Danielson, A. L. R. et al.  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  emission from a dense gas disc at  $z = 2.3$ : abundance variations, cosmic rays and the initial conditions for star formation. *Mon. Not. R. Astron. Soc.* **436**, 2793–2809 (2013).
- Barnes, P. J. et al. The three-mm ultimate Mopra Milky Way Survey. I. Survey overview, initial data releases, and first results. *Astrophys. J.* **812**, 6 (2015).
- Jiménez-Donaire, M. J. et al.  $^{13}\text{CO}/\text{C}^{18}\text{O}$  gradients across the disks of nearby spiral galaxies. *Astrophys. J.* **836**, L29 (2017).
- Ballero, S. K., Matteucci, F., Origlia, L. & Rich, R. M. Formation and evolution of the Galactic bulge: constraints from stellar abundances. *Astron. Astrophys.* **467**, 123–136 (2007).
- Dabringhausen, J., Kroupa, P. & Baumgardt, H. A top-heavy stellar initial mass function in starbursts as an explanation for the high mass-to-light ratios of ultra-compact dwarf galaxies. *Mon. Not. R. Astron. Soc.* **394**, 1529–1543 (2009).



18. Dabringhausen, J., Kroupa, P., Pflamm-Altenburg, J. & Mieske, S. Low-mass X-ray binaries indicate a top-heavy stellar initial mass function in ultracompact dwarf galaxies. *Astrophys. J.* **747**, 72 (2012).
19. Peacock, M. B. et al. Further constraints on variations in the initial mass function from low-mass X-ray binary populations. *Astrophys. J.* **841**, 28 (2017).
20. Schneider, F. R. N. et al. An excess of massive stars in the local 30 Doradus starburst. *J. Sci.* **359**, 69–71 (2018).
21. Banerjee, S. & Kroupa, P. On the true shape of the upper end of the stellar initial mass function. The case of R136. *Astron. Astrophys.* **547**, A23 (2012).
22. Kalari, V. M., Carraro, G., Evans, C. J. & Rubio, M. The Magellanic Bridge cluster NGC 796: deep optical AO imaging reveals the stellar content and initial mass function of a massive open cluster. *Astrophys. J.* **857**, 132 (2018).
23. Lee, J. C. et al. Comparison of H $\alpha$  and UV star formation rates in the local volume: systematic discrepancies for dwarf galaxies. *Astrophys. J.* **706**, 599–613 (2009).
24. Pflamm-Altenburg, J. & Kroupa, P. Clustered star formation as a natural explanation for the H $\alpha$  cut-off in disk galaxies. *Nature* **455**, 641–643 (2008).
25. Speagle, J. S., Steinhardt, C. L., Capak, P. L. & Silverman, J. D. A highly consistent framework for the evolution of the star-forming “main sequence” from  $z \sim 0$ –6. *Astrophys. J. Suppl. Ser.* **214**, 15 (2014).
26. Madau, P. et al. High-redshift galaxies in the Hubble deep field: colour selection and star formation history to  $z \sim 4$ . *Mon. Not. R. Astron. Soc.* **283**, 1388–1404 (1996).
27. Pflamm-Altenburg, J. & Kroupa, P. The fundamental gas depletion and stellar-mass buildup times of star-forming galaxies. *Astrophys. J.* **706**, 516–524 (2009).
28. Heikkilä, A., Johansson, L. E. B. & Olofsson, H. The C<sup>18</sup>O/C<sup>17</sup>O ratio in the Large Magellanic Cloud. *Astron. Astrophys.* **332**, 493–502 (1998).
29. Muraoka, K. et al. ALMA Observations of N83C in the early stage of star formation in the Small Magellanic Cloud. *Astrophys. J.* **844**, 98 (2017).
30. Nishimura, Y. et al. Spectral line survey toward a molecular cloud in IC10. *Astrophys. J.* **829**, 94 (2016).

**Acknowledgements** Z.-Y.Z. is grateful to X. Fu, H.-Y. B. Liu, Y. Shirley and P. Barnes for discussions. Z.-Y.Z., R.J.I. and P.P.P. acknowledge support from

the European Research Council in the form of the Advanced Investigator Programme, 321302, COSMICISM. F.M. acknowledges financial funds from Trieste University, FRA2016. This research was supported by the Munich Institute for Astro- and Particle Physics (MIAPP) of the DFG cluster of excellence “Origin and Structure of the Universe”. This work also benefited from the International Space Science Institute (ISSI) in Bern, thanks to the funding of the team “The Formation and Evolution of the Galactic Halo” (Principal Investigator D.R.) This paper makes use of the ALMA data. ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (South Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ.

**Reviewer information** *Nature* thanks C. Henkel, P. Kroupa and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** Z.-Y.Z. is the Principal Investigator of the ALMA observing project. Z.-Y.Z. reduced the data and wrote the initial manuscript. R.J.I. and P.P.P. provided ideas to initialize the project and helped write the manuscript. Z.-Y.Z. and P.P.P. worked on molecular line modeling of isotopologue ratios and chemical/thermal effects on the abundances. D.R. and F.M. ran the chemical evolution models and provided theoretical interpretation of the data. All authors discussed and commented on the manuscript.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0196-x>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to R.J.I.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

**Sample.** Our sample comprises the strongest CO emitters in the early Universe, taken from the literature<sup>31–33</sup>: four strongly lensed submillimetre galaxies at  $z \approx 2–3$ , with look-back times exceeding ten billion years. Two of these galaxies, SPT-S J010312–4538.8 ( $z = 3.09$ , also known as SPT 0103–45) and SPT-S J012506–4723.7 ( $z = 2.51$ , also known as SPT 0125–47), were selected<sup>32</sup> using the South Pole Telescope at wavelengths  $\lambda = 1.4$  mm and 2 mm; another, H-ATLAS J090302.9–014127 ( $z = 2.31$ , also known as SDP.17b), was discovered using the Herschel Space Observatory<sup>34</sup> at far-infrared wavelengths; the last, H1413+117 ( $z = 2.56$ , the ‘Cloverleaf’ quasar), was discovered as a result of its rare quadruple-spot optical morphology, and was later found to be bright in CO and in the dust continuum<sup>31,35</sup>. We list the basic characteristics of the sample in Extended Data Table 1.

**Observations and data reduction.** We have performed simultaneous observations of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  using ALMA in its relatively compact array configurations (C36-1 and C36-2), with two 2-GHz-wide spectral windows in bands 3 and 4. We used the remaining two spectral windows to cover continuum emission. Between 10 min and 30 min were spent on target for each transition. For SDP.17b, we observed both the  $J = 3 \rightarrow 2$  and  $J = 4 \rightarrow 3$  transitions of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , in order to have a redundant measurement for their line ratios as well as constraints on the relative excitation in these rare isotopic lines. We also observed  $^{12}\text{CO } J = 5 \rightarrow 4$  for SPT 0103–45 and  $^{12}\text{CO } J = 4 \rightarrow 3$  for SPT 0125–47, with similar configurations. Calibrators, integration time and atmospheric conditions are listed in Extended Data Table 2.

All the data were calibrated manually using CASA v4.7.1<sup>36</sup>, using standard procedures. We subtracted the continuum using the CASA task, uvcontsub, by fitting a linear slope to the line-free channels. We cleaned the visibility data with a channel width of about 20–30 km s<sup>−1</sup>, using a Briggs weighting with robust = 1.5 to optimize sensitivity. We applied a primary beam correction to all the cleaned data. Our target galaxies are mostly unresolved, or only marginally resolved. We assume that linewidths of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  are the same as those of  $^{12}\text{CO}$  lines, to minimize uncertainties in the line flux fitting. Extended Data Figs. 1–3 present the velocity-integrated flux (moment 0) maps of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , overlaid with contours of high-resolution submillimetre continuum. In Extended Data Fig. 4 we present the spectra. SPT 0103–45 has two velocity components that cover a very large velocity span. The overall line profile of  $^{13}\text{CO}$  is consistent with  $^{12}\text{CO}$ , but limited by the noise level. We adopt only the narrow (stronger) component, seen for the yellow shadow region in the  $^{12}\text{CO } J = 5 \rightarrow 4$  spectrum, to avoid confusion from the broad (weaker) component (see Extended Data Fig. 4). Our synthesized beam sizes are mostly larger than, or at least comparable to, the apparent sizes revealed by the high-resolution submillimetre continuum images, so any missing flux is expected to be negligible. We extract spectra of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  using circular apertures equal to 4″–6″ in diameter, as shown in Extended Data Figs. 1–3.

To measure velocity-integrated line fluxes and the associated errors, we performed three independent methods: we first fitted one-dimensional Gaussian profiles to the extracted spectra with a fixed linewidth from  $^{12}\text{CO}$ , and fixed the frequency interval between  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , since we can assume confidently that these two lines are emitted from the same excitation component, such that their line centres do not shift relative to one another. Second, we made moment-0 maps and fitted two-dimensional Gaussian profiles, as an independent check of line flux. Third, to better estimate the noise level, we also calculated the theoretical noise using the ALMA sensitivity calculator (<https://almascience.eso.org/proposing/sensitivity-calculator>), given the integration time, precipitable water vapour, linewidth and array configuration. The measured line fluxes and properties are listed in Extended Data Table 3. To be conservative, in Figs. 1 and 2 we adopt the largest error among the results from the three methods in the analysis.

**Line ratios and optical depths.** *Conditions of LTE.* We first analyse the observed line ratios of  $^{13}\text{CO}$  to  $\text{C}^{18}\text{O}$ , to constrain the molecular line optical depths, assuming LTE ( $T_{\text{ex}}^{\text{line}} = T_{\text{kin}}^{\text{line}}$ ). The line brightness temperature ratios of  $^{12}\text{CO}$  to  $^{13}\text{CO}$  and  $^{13}\text{CO}$  to  $\text{C}^{18}\text{O}$  can be expressed as:

$$\frac{T_{\text{b}}^{12\text{CO}}}{T_{\text{b}}^{13\text{CO}}} = \frac{J_{\nu}(T_{\text{ex}}^{12\text{CO}}) - J_{\nu}(T_{\text{bg}}^{12\text{CO}})}{J_{\nu}(T_{\text{ex}}^{13\text{CO}}) - J_{\nu}(T_{\text{bg}}^{13\text{CO}})} \times \frac{1 - \exp(-\tau^{12\text{CO}})}{1 - \exp(-\tau^{13\text{CO}})} \quad (1)$$

and

$$\frac{T_{\text{b}}^{13\text{CO}}}{T_{\text{b}}^{\text{C}^{18}\text{O}}} = \frac{J_{\nu}(T_{\text{ex}}^{13\text{CO}}) - J_{\nu}(T_{\text{bg}}^{13\text{CO}})}{J_{\nu}(T_{\text{ex}}^{\text{C}^{18}\text{O}}) - J_{\nu}(T_{\text{bg}}^{\text{C}^{18}\text{O}})} \times \frac{1 - \exp(-\tau^{13\text{CO}})}{1 - \exp(-\tau^{\text{C}^{18}\text{O}})} \quad (2)$$

where  $T_{\text{b}}$  is the brightness temperature of the molecular transition,  $T_{\text{ex}}$  is the excitation temperature, and  $T_{\text{bg}}$  is the radiation temperature of the background emission field, which is dominated by the cosmic microwave background, following  $T_{\text{CMB}} \approx 2.73 \times (1+z)$  K, where  $z$  is the redshift.  $\tau^{\text{line}}$  is the optical depth of the given

transition.  $J_{\nu}(T) = (h\nu/k_{\text{B}})/\{\exp[h\nu/(k_{\text{B}}T)] - 1\}$  is the Planck radiation temperature at the rest frequency of the line emission,  $\nu^{\text{line}}$ .  $k_{\text{B}}$  is the Boltzmann constant,  $h$  is the Planck constant, and  $T$  is the temperature considered. For optically thick lines (for example,  $^{12}\text{CO}$  for most conditions),  $1 - \exp(-\tau^{\text{line}}) \approx 1$ ; for optically thin lines (for example,  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ ),  $1 - \exp(-\tau^{\text{line}}) \approx \tau^{\text{line}}$ .

In Extended Data Fig. 5 we present the  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  and  $I(^{12}\text{CO})/I(^{13}\text{CO})$  velocity-integrated line intensity ratios as a function of the optical depths of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , under LTE conditions. We also calculated the corresponding  $\text{H}_2$  column densities, assuming Galactic abundances<sup>37,38</sup>. A representative Galactic abundance ratio of  $^{12}\text{CO}/^{13}\text{CO} = 70$  is assumed, and the optical depth for  $^{13}\text{CO}$  needs to be  $< 0.03$ , which is around  $10 \times$  lower than the Galactic average values<sup>14</sup> for producing the observed high  $I(^{12}\text{CO})/I(^{13}\text{CO})$  ratios,  $\geq 30$ . The corresponding optical depth of  $^{12}\text{CO}$  is about 2, much lower than the typical values for  $^{12}\text{CO } J = 1 \rightarrow 0$  found in typical Galactic molecular clouds<sup>14</sup>, but consistent with a moderate optical depth of  $^{12}\text{CO}$  found in local starburst galaxies<sup>39</sup>.

Only when the optical depth of  $\text{C}^{18}\text{O}$  is much greater than unity (corresponding to  $\tau^{13\text{CO}}$  being much greater than 7, which makes the molecular hydrogen column density  $N_{\text{H}_2}$  much greater than  $10^{25} \text{ cm}^{-2}$ ), does the  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  line ratio approach unity. In this case the line ratios of  $I(^{12}\text{CO})/I(^{13}\text{CO})$  and  $I(^{12}\text{CO})/I(\text{C}^{18}\text{O})$  would also move towards unity, in conflict with our observed ratios. Even for moderate values of  $\tau^{13\text{CO}} \approx 0.2–0.5$ , the line ratio of  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  stays at about 6–7, not strongly biased by  $\tau^{13\text{CO}}$ .

*Non-LTE conditions.* In Extended Data Fig. 6, we present non-LTE models derived with a non-LTE radiative transfer code, RADEX<sup>40</sup>, showing the optical depth,  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  and  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratios as a function of  $^{13}\text{CO}$  column density and molecular gas column density,  $N_{\text{H}_2}$ . We calculate for different  $\text{H}_2$  volume densities,  $n_{\text{H}_2}$ , of  $10^3 \text{ cm}^{-3}$ ,  $10^4 \text{ cm}^{-3}$  and  $10^5 \text{ cm}^{-3}$ , covering the most common  $n_{\text{H}_2}$  range—typical values from normal molecular clouds to dense cores. We assume the same abundance ratios as we assumed for the LTE conditions, that is,  $^{12}\text{CO}/^{13}\text{CO} = 70$  and  $^{13}\text{CO}/\text{C}^{18}\text{O} = 7$ , which are values representative of the Milky Way disk. The velocity width (full-width at half-maximum, FWHM) is set to  $300 \text{ km s}^{-1}$ , as the typical (indeed, at the lower end) linewidth found in ULIRGs and submillimetre galaxies<sup>41</sup>. In Extended Data Fig. 6b and c we overlay the LTE results, for comparison.

For all models, we set the kinetic temperature,  $T_{\text{kin}}$ , to be 30 K, which is a typical dust temperature for the submillimetre galaxy population<sup>42</sup>, and is also the lower limit of the kinetic temperature of the  $\text{H}_2$  gas, as the minimum temperature powered by the cosmic-ray heating for such starburst conditions<sup>43</sup>. Higher  $T_{\text{kin}}$  would bring the CO energy population towards higher- $J$  transitions, making optical depths even smaller.

Extended Data Fig. 6 shows that, for  $n_{\text{H}_2} = 10^3 \text{ cm}^{-3}$ , which is a typical value for normal Galactic molecular cloud conditions, only when  $N_{\text{H}_2}$  is much greater than  $10^{26} \text{ cm}^{-2}$ , the line ratio of  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  can approach unity ( $< 1.5$ , considering the uncertainties in line ratios). The required high column densities are a few orders of magnitude higher than the typical values measured in submillimetre galaxies: about  $10^{23}–10^{24} \text{ cm}^{-2}$ , obtained using X-rays<sup>44</sup>, CO radiative transfer modelling<sup>45</sup>, and dust<sup>42</sup>. This is especially supported by the Cloverleaf quasar, whose X-ray emission has been clearly detected<sup>46</sup>, given the Compton limit of about  $10^{24} \text{ cm}^{-2}$ .

On the other hand, the high-density results, that is, for  $n_{\text{H}_2} = 10^4 \text{ cm}^{-3}$  and  $10^5 \text{ cm}^{-3}$ , are very similar to those under LTE conditions. For all conditions, a moderate  $^{13}\text{CO}$  optical depth does not vary the  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  ratio much from the abundance ratio. So, it is highly unlikely that the unity value of the  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  ratio can be caused by high optical depths.

**Possible HNCO contamination of  $\text{C}^{18}\text{O}$  lines.** HNCO  $5_{05} \rightarrow 4_{04}$  ( $J = 5 \rightarrow 4$ ) has a rest frequency of 109.9058 GHz, close to the rest frequency of  $\text{C}^{18}\text{O } J = 1 \rightarrow 0$  (109.7822 GHz), with a velocity offset of about  $370 \text{ km s}^{-1}$ . When the linewidths are broad, these two lines are sometimes blended, which leads to a possible contamination of the  $\text{C}^{18}\text{O}$  measurements<sup>12,47,48</sup>. In these observations it has been found that HNCO  $J = 5 \rightarrow 4$  could contribute up to about 30% of the total flux ( $\text{C}^{18}\text{O} + \text{HNCO}$ ), for the most extreme cases in the local Universe, for example, Arp 220<sup>48</sup> and IRAS 13120–5453<sup>12</sup>.

HNCO  $J = 5 \rightarrow 4$  has a critical density,  $n_{\text{crit}} \approx 10^6 \text{ cm}^{-3}$  and is regarded as a dense-gas tracer that could be excited in slow-shock regions<sup>49,50</sup>. The Einstein A coefficient increases as  $A \propto (J+1)^3$ , so  $n_{\text{crit}}$  increases quickly for high- $J$  transitions, such as HNCO  $J = 15 \rightarrow 14$  ( $\nu_{\text{rest}} = 329.66 \text{ GHz}$ ) and  $J = 20 \rightarrow 19$  ( $\nu_{\text{rest}} = 439.62 \text{ GHz}$ ), contaminating the  $\text{C}^{18}\text{O } J = 3 \rightarrow 2$  and  $J = 4 \rightarrow 3$  transitions involved in our study much less.

To estimate better how much HNCO lines may contaminate the  $\text{C}^{18}\text{O}$  lines, in Extended Data Fig. 7 we show the theoretical line ratios between HNCO and  $\text{C}^{18}\text{O}$  and high- $J$   $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  ratios normalized with  $I(\text{HNCO } J = 5 \rightarrow 4)/I(\text{C}^{18}\text{O } J = 1 \rightarrow 0)$ , using RADEX<sup>40</sup>. We assume the same abundances measured in Arp 220<sup>47</sup>, and use molecular data from the Leiden Atomic and Molecular Database<sup>51</sup> (LAMDA). We assume  $T_{\text{kin}} = 30 \text{ K}$  to be the representative temperature of the  $\text{H}_2$  gas.

Extended Data Fig. 7a shows that the  $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  ratio increases with  $n_{\text{H}_2}$ . Moreover, the ratio decreases quickly with  $J$  transitions, meaning that the contamination from HNCO to  $\text{C}^{18}\text{O}$  is much less severe for the high- $J$  transitions, compared to that of the  $\text{C}^{18}\text{O } J=1 \rightarrow 0$  line. Extended Data Fig. 7b shows  $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  line brightness ratios normalized by  $I(\text{HNCO } J=5 \rightarrow 4)/I(\text{C}^{18}\text{O } J=1 \rightarrow 0)$ . With a weak dependency on  $n_{\text{H}_2}$ , the ratios of  $I(\text{HNCO } J=10 \rightarrow 9)/I(\text{HNCO } J=5 \rightarrow 4)$  and  $I(\text{HNCO } J=15 \rightarrow 14)/I(\text{HNCO } J=5 \rightarrow 4)$  are one order of magnitude lower than unity. Unfortunately the LAMBDA database does not have the data for the transition of  $\text{HNCO } J=20 \rightarrow 19$ , whose HNCO/ $\text{C}^{18}\text{O}$  ratio is expected to be even lower.

Even if we take the highest  $\text{HNCO } J=5 \rightarrow 4$  contamination found in local galaxies, that is, 30%, the corresponding contamination for the high- $J\text{C}^{18}\text{O}$  transitions will be at most 3%, which can be regarded as negligible for the lines in our study.

**Chemical evolution model.** We adopt a single-zone chemical evolution model for our analysis, originally developed to describe the evolution of the Milky Way<sup>52</sup>, then further extended to other galaxies<sup>53</sup>. The model computes the evolution of abundances of multiple elements, including  $^{12}\text{C}$ ,  $^{16}\text{O}$ ,  $^{13}\text{C}$  and  $^{18}\text{O}$  in the ISM of galaxies. We use detailed numerical models to solve the classical set of equations of chemical evolution<sup>10,52–56</sup>, with the following assumptions:

(1) Gas inflow with primordial chemical composition provides raw material for star formation. The gas is accreted at an exponentially fading rate and the timescale of the process is a free parameter of the model.

(2) Galactic outflows remove both the stellar ejecta and a fraction of the ambient ISM.

(3) Star formation follows the canonical Kennicutt–Schmidt law<sup>57</sup>; the masses of the newly-formed stars follow the input IMF.

(4) Finite stellar lifetimes for different stars need to be considered (that is, no instantaneous recycling approximation (non-IRA) is adopted)<sup>58</sup>.

(5) Stars release the elements they have synthesized during their lifetime, as well as those already present when they were born that are left unaltered by the nucleosynthesis processes, when they die.

(6) Stellar ejecta are mixed with the ISM homogeneously.

The yields adopted account for the dependence of several stellar processes on the initial metallicity of the stars, and have been calibrated with the best fit using the Milky Way data, which are relevant to a range of metallicity and evolution timescales<sup>6</sup>. The time-delay effect is considered in the chemical evolution, namely the differences between the lifetimes of massive stars and low-mass stars<sup>59</sup>. We used an analytical formula for the stellar lifetimes that linearly interpolate stellar lifetime tables<sup>58</sup>. The time lag in producing and releasing primary (those synthesized directly from H and He; that is,  $^{12}\text{C}$  and  $^{16}\text{O}$ ) and secondary elements (those derived from metals already present in the star at birth; that is,  $^{13}\text{C}$  and  $^{18}\text{O}$ )—but note that a fraction of  $^{13}\text{C}$  is also synthesized as a primary element—is also considered<sup>52</sup>. These two effects, corroborated by the star-formation history and the IMF adopted, determine the amount of different isotopes released to the ISM on different timescales. In particular, the bulk of  $^{13}\text{C}$  is released later than  $^{12}\text{C}$ , and the bulk of  $^{18}\text{O}$  is released later than  $^{16}\text{O}$ . Chemical evolution models can now follow the evolution of various isotopic ratios, tracing abundance ratios not only between the isotopes of each element<sup>6</sup>, but also between different elements.

The most important aspect regarding this work is that such models can now compare the effects of a young starburst (with a regular stellar IMF, for example, the Kroupa IMF) against those due to different stellar IMFs via carefully chosen isotope (and thus isotopologue) ratios. This critical advance was made by no longer assuming instantaneous element enrichment of the ISM by the stars, but incorporating the different timescales of their release into the ISM. It should be noted that these timescales, and the relative delays between the release of various isotopes into the ISM, are set by stellar physics, that is, they are not free parameters. With the Kroupa IMF, only an unphysical combination of star-formation history, that is, for a starburst timescale  $\tau < 10$  Myr and a star-formation rate of more than  $20,000 M_{\odot} \text{ yr}^{-1}$ , could approach the observed  $^{13}\text{C}/^{18}\text{O}$  ratios of near unity.

**Origins of carbon and oxygen isotopes.** The  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratios have been found to vary systematically in galaxies with different star-formation rates and Hubble types<sup>39,61</sup>. Owing to the differences in the origins of  $^{12}\text{C}$  and  $^{13}\text{C}$ , it has been proposed that the  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratios can be used to derive their abundance ratio, which can further probe the stellar IMF, or different star-formation modes<sup>12,43,60,62</sup>.

The  $^{12}\text{C}$  element is primarily produced by helium burning (the classical triple- $\alpha$  process), and multiple channels can produce  $^{12}\text{C}$  in nucleosynthesis<sup>63</sup>. In the Milky Way,  $^{12}\text{C}$  is primarily produced by low- and intermediate-mass stars, revealed by data for stars in the solar vicinity: in fact, the  $[\text{C}/\text{Fe}]$  ratio as a function of  $[\text{Fe}/\text{H}]$  is almost constant, with  $[\text{C}/\text{Fe}]$  being solar, indicating that C and Fe are produced in the same proportions by the same stars<sup>9,64</sup>. If mass loss from massive stars is considered,  $^{12}\text{C}$  released from massive stars still accounts for  $< 50\%$  of the total<sup>65,66</sup>. On the other hand,  $^{13}\text{C}$  is released mainly from low- and intermediate-mass stars largely as a secondary element, because  $^{13}\text{C}$  production needs a pre-existing seed, namely,

the primary element,  $^{12}\text{C}$ <sup>63,67</sup>.  $^{13}\text{C}$  also has a primary component in nucleosynthesis but it can only occur in red asymptotic giant branch stars, where periodic dredge-up episodes convect  $^{12}\text{C}$  to the stellar surface and form  $^{13}\text{C}$ .

Since both  $^{12}\text{C}$  and  $^{13}\text{C}$  are mostly produced by low- and intermediate-mass stars, the  $^{12}\text{C}/^{13}\text{C}$  ratio cannot discriminate between IMFs unambiguously. In previous work, we show that by switching from the Ballero IMF<sup>16</sup> (a very top-heavy IMF, which can reproduce the chemical abundances of stars in the Galactic bulge) to the Kroupa IMF, the  $^{12}\text{C}/^{13}\text{C}$  ratio varies<sup>6</sup> only by a factor of 2, indicating that carbon isotopologue ratios are not very sensitive to IMF. Furthermore, the  $^{12}\text{C}$  abundance is very difficult to obtain because the  $^{12}\text{C}$ -bearing major isotopologue lines are mostly optically thick.

The origin of oxygen isotopes is rather different. As earlier work suggested<sup>68</sup>, the stellar yields of  $^{16}\text{O}$ , and  $^{18}\text{O}$  are sensitive to different stellar masses, owing to their temperature sensitivity in stellar nucleosynthesis<sup>69</sup>. Production of  $^{16}\text{O}$  is dominated by massive stars<sup>54</sup>, as revealed by chemical evolution models in the Milky Way using detailed stellar yields<sup>6</sup>. Only a tiny fraction of  $^{16}\text{O}$  is contributed by asymptotic giant branch stars<sup>67</sup>. Massive stars also dominate the production<sup>70</sup> of  $^{18}\text{O}$ , predominantly in the early stages of helium burning<sup>69</sup>. As a secondary element, the  $^{18}\text{O}$  yield relies strongly on the pre-existence of  $^{16}\text{O}$ , so the metallicity in oxygen also plays a major part in producing  $^{18}\text{O}$ . The production of  $^{18}\text{O}$  is biased to more massive stars compared to  $^{13}\text{C}$  which is more biased to low- and intermediate-mass stars. So, the abundance ratio of  $^{13}\text{C}$  and  $^{18}\text{O}$  does indeed reflect different IMFs (see Fig. 2).

The abundance ratios of  $^{12}\text{C}/^{13}\text{C}$  and  $^{16}\text{O}/^{18}\text{O}$  can trace star-formation timing and IMF<sup>6</sup>, respectively. The  $^{12}\text{C}$  and  $^{16}\text{O}$  abundances are compromised by the optical depths of molecular lines, which are difficult to estimate accurately. However, the combination of both carbon and oxygen isotopologues—the abundance ratio of  $^{13}\text{C}$  and  $^{18}\text{O}$ —can be obtained easily from the observed intensity ratio of two optically thin lines,  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , in the same  $J$  transition. Moreover, these two lines can be obtained simultaneously using current facilities, owing to the close spacing of their rest frequencies. They have almost identical critical densities and upper energy levels, essentially free from excitation differences. Even for strongly lensed galaxies, it is safe to assume that any differential lensing effect between the two lines is negligible.

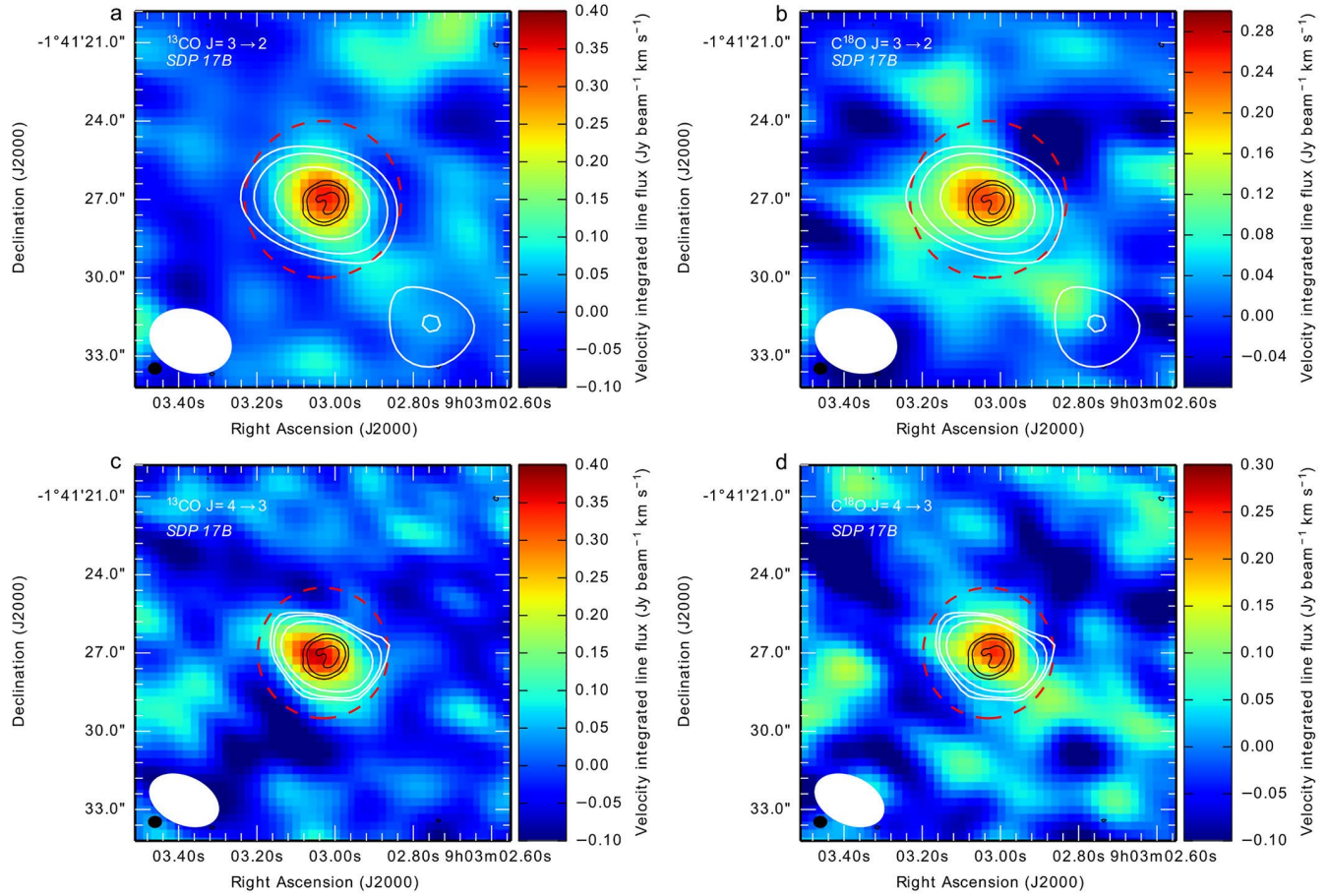
**Code availability.** We opt not to make the code used for the chemical evolution modelling publicly available because it is an important asset of the researchers' toolkits. The code for analysing the line ratios and optical depths of  $^{12}\text{CO}$ ,  $^{13}\text{CO}$ ,  $\text{C}^{18}\text{O}$  are based on the publicly available non-LTE radiative transfer code, RADEX (<https://personal.sron.nl/~vdtak/radex/index.shtml>).

**Data availability.** The dataset that supports the findings of this study is available in the ALMA archive (<http://almascience.eso.org/aq/>) under the observing projects 2015.1.01309.S (<http://almascience.eso.org/aq/?projectcode=2015.1.01309.S>), 2013.1.00164.S (<http://almascience.eso.org/aq/?projectcode=2013.1.00164.S>), 2011.0.00958.S (<http://almascience.eso.org/aq/?projectcode=2011.0.00958.S>) and 2011.0.00747.S (<http://almascience.eso.org/aq/?projectcode=2011.0.00747.S>). Additional requests can be directed to the corresponding author.

- Magain, P., Surdej, J., Swings, J.-P., Borgeest, U. & Kayser, R. Discovery of a quadruply lensed quasar—the 'clover leaf' H1413 + 117. *Nature* **334**, 325–327 (1988).
- Weiβ, A. et al. ALMA redshifts of millimeter-selected galaxies from the SPT Survey: the redshift distribution of dusty star-forming galaxies. *Astrophys. J.* **767**, 88 (2013).
- Negrello, M. et al. The detection of a population of submillimeter-bright, strongly lensed galaxies. *J. Sci.* **330**, 800 (2010).
- Griffin, M. J. et al. The Herschel-SPIRE instrument and its in-flight performance. *Astron. Astrophys.* **518**, L3 (2010).
- Solomon, P., Vanden Bout, P., Carilli, C. & Guelin, M. The essential signature of a massive starburst in a distant quasar. *Nature* **426**, 636–638 (2003).
- McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. In *Astronomical Data Analysis Software and Systems XVI* (eds Shaw, R. A., Hill, F. & Bell, D. J.) Vol. 376, 127 (Astronomical Society of the Pacific Conference Series, ASP, 2007).
- Mangum, J. G. & Shirley, Y. L. How to calculate molecular column density. *Publ. Astron. Soc. Pacif.* **127**, 266 (2015).
- Frerking, M. A., Langer, W. D. & Wilson, R. W. The relationship between carbon monoxide abundance and visual extinction in interstellar clouds. *Astrophys. J.* **262**, 590–605 (1982).
- Aalto, S., Booth, R. S., Black, J. H. & Johansson, L. E. B. Molecular gas in starburst galaxies: line intensities and physical conditions. *Astron. Astrophys.* **300**, 369 (1995).
- van der Tak, F. F. S., Black, J. H., Schöier, F. L., Jansen, D. J. & van Dishoeck, E. F. A computer program for fast non-LTE analysis of interstellar line spectra. With diagnostic plots to interpret observed line intensity ratios. *Astron. Astrophys.* **468**, 627–635 (2007).
- Yang, C. et al. Molecular gas in the Herschel-selected strongly lensed submillimeter galaxies at  $z \sim 2-4$  as probed by multi-J CO lines. *Astron. Astrophys.* **608**, A144 (2017).
- Simpson, J. M. et al. The SCUBA-2 Cosmology Legacy Survey: multi-wavelength properties of ALMA-identified submillimeter galaxies in UKIDSS UDS. *Astrophys. J.* **839**, 58 (2017).

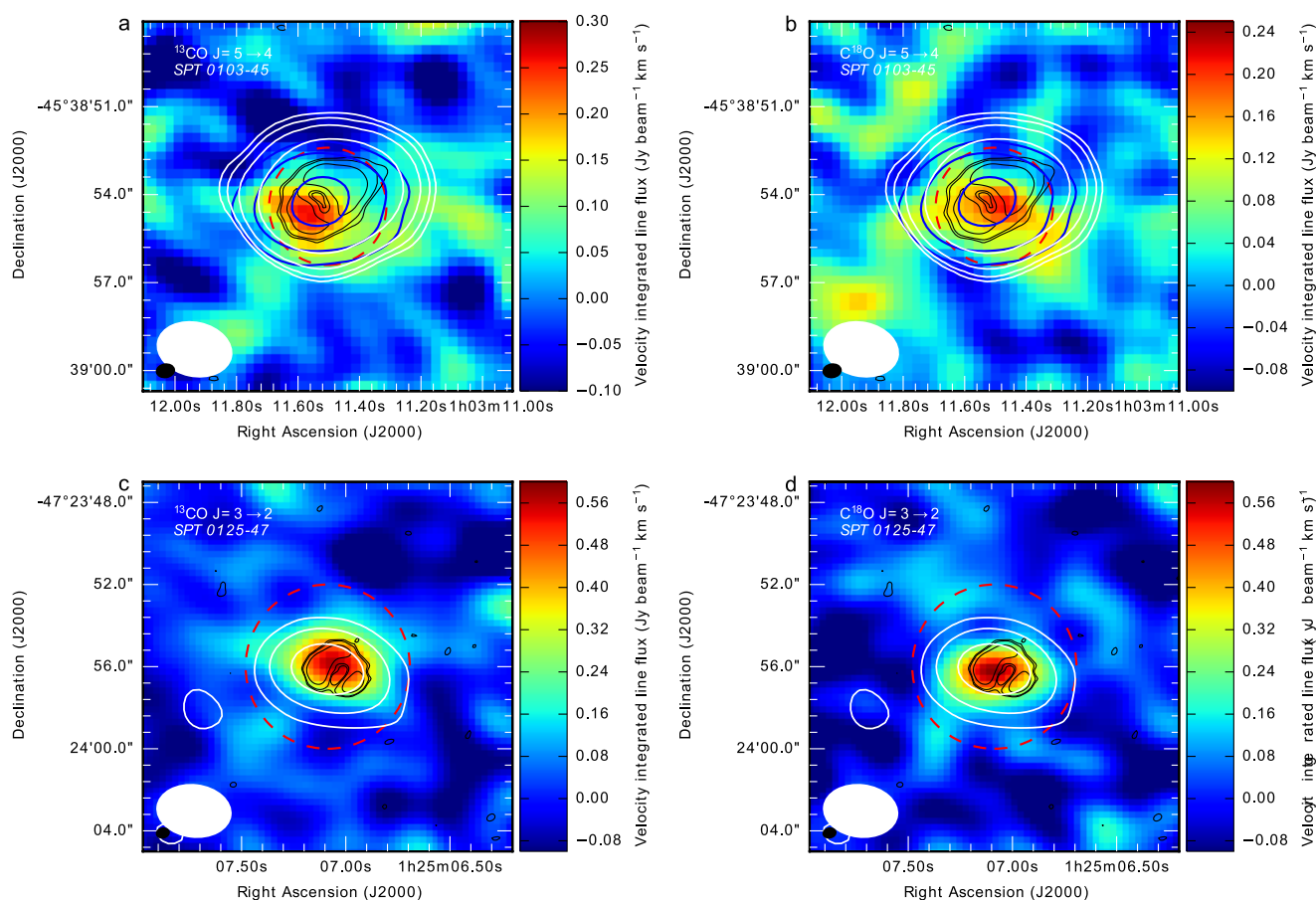


43. Papadopoulos, P. P. et al. Molecular gas heating mechanisms, and star formation feedback in merger/starbursts: NGC 6240 and Arp 193 as case studies. *Astrophys. J.* **788**, 153 (2014).
44. Wang, S. X. et al. An ALMA survey of submillimeter galaxies in the extended Chandra deep field-south: the AGN fraction and X-ray properties of submillimeter galaxies. *Astrophys. J.* **778**, 179 (2013).
45. Spilker, J. S. et al. The rest-frame submillimeter spectrum of high-redshift, dusty, star-forming galaxies. *Astrophys. J.* **785**, 149 (2014).
46. Chartas, G., Eracleous, M., Agol, E. & Gallagher, S. C. Chandra observations of the Cloverleaf quasar H1413+117: a unique laboratory for microlensing studies of a LoBAL quasar. *Astrophys. J.* **606**, 78–84 (2004).
47. Martín, S., Martín-Pintado, J. & Mauersberger, R. HNC/O abundances in galaxies: tracing the evolutionary state of starbursts. *Astrophys. J.* **694**, 610–617 (2009).
48. Greve, T. R., Papadopoulos, P. P., Gao, Y. & Radford, S. J. E. Molecular gas in extreme star-forming environments: the starbursts Arp 220 and NGC 6240 as case studies. *Astrophys. J.* **692**, 1432–1446 (2009).
49. Zinchenko, I., Henkel, C. & Mao, R. Q. HNC/O in massive galactic dense cores. *Astron. Astrophys.* **361**, 1079–1094 (2000).
50. Li, J., Wang, J. Z., Gu, Q. S. & Zheng, X. W. Distribution of HNC/O  $5_{05}-4_{04}$  in massive star-forming regions. *Astron. Astrophys.* **555**, A18 (2013).
51. Schöier, F. L., van der Tak, F. F. S., van Dishoeck, E. F. & Black, J. H. An atomic and molecular database for analysis of submillimetre line observations. *Astron. Astrophys.* **432**, 369 (2005).
52. Matteucci, F. *Chemical Evolution of Galaxies* (Springer, Berlin, 2012).
53. Romano, D., Bellazzini, M., Starkenburg, E. & Leaman, R. Chemical enrichment in very low metallicity environments: Boötes I. *Mon. Not. R. Astron. Soc.* **446**, 4220–4231 (2015).
54. Tinsley, B. M. Evolution of the stars and gas in galaxies. *Fundamentals Cosm. Phys.* **5**, 287–388 (1980).
55. Pagel, B. E. J. *Nucleosynthesis and Chemical Evolution of Galaxies* (Cambridge Univ. Press, Cambridge, 1997).
56. Matteucci, F. (ed.) *The Chemical Evolution of the Galaxy* Vol. 253 (Springer, Netherlands, 2001).
57. Kennicutt, R. C. Jr. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498**, 541–552 (1998).
58. Schaller, G., Schaerer, D., Meynet, G. & Maeder, A. New grids of stellar models from 0.8 to 120 solar masses at  $Z = 0.020$  and  $Z = 0.001$ . *Astron. Astrophys. Suppl.* **96**, 269–331 (1992).
59. Matteucci, F. & Greggio, L. Relative roles of type I and II supernovae in the chemical enrichment of the interstellar gas. *Astron. Astrophys.* **154**, 279–287 (1986).
60. Henkel, C. & Mauersberger, R. C and O nucleosynthesis in starbursts - the connection between distant mergers, the Galaxy and the solar system. *Astron. Astrophys.* **274**, 730–742 (1993).
61. Davis, T. A. Systematic variation of the  $^{12}\text{CO}/^{13}\text{CO}$  ratio as a function of star formation rate surface density. *Mon. Not. R. Astron. Soc.* **445**, 2378–2384 (2014).
62. Henkel, C., Downes, D., Weiß, A., Riechers, D. & Walter, F. Weak  $^{13}\text{CO}$  in the Cloverleaf quasar: evidence for a young, early generation starburst. *Astron. Astrophys.* **516**, A111 (2010).
63. Hughes, G. L. et al. The evolution of carbon, sulphur and titanium isotopes from high redshift to the local Universe. *Mon. Not. R. Astron. Soc.* **390**, 1710–1718 (2008).
64. Nomoto, K., Tominaga, N., Umeda, H., Kobayashi, C. & Maeda, K. Nucleosynthesis yields of core-collapse supernovae and hypernovae, and galactic chemical evolution. *Nucl. Phys. A* **777**, 424–458 (2006).
65. Cescutti, G., Matteucci, F., McWilliam, A. & Chiappini, C. The evolution of carbon and oxygen in the bulge and disk of the Milky Way. *Astron. Astrophys.* **505**, 605–612 (2009).
66. Carigi, L., Peimbert, M., Esteban, C. & Garca-Rojas, J. Carbon, nitrogen, and oxygen galactic gradients: a solution to the carbon enrichment problem. *Astrophys. J.* **623**, 213–224 (2005).
67. Meyer, B. S., Nittler, L. R., Nguyen, A. N. & Messenger, S. Nucleosynthesis and chemical evolution of oxygen. *Rev. Mineral. Geochem.* **68**, 31–53 (2008).
68. Sage, L. J., Henkel, C. & Mauersberger, R. Extragalactic O-18/O-17 ratios and star formation—high-mass stars preferred in starburst systems? *Astron. Astrophys.* **249**, 31–35 (1991).
69. Kobayashi, C., Karakas, A. I. & Umeda, H. The evolution of isotope ratios in the Milky Way Galaxy. *Mon. Not. R. Astron. Soc.* **414**, 3231–3250 (2011).
70. Timmes, F. X., Woosley, S. E. & Weaver, T. A. Galactic chemical evolution: hydrogen through zinc. *Astrophys. J. Suppl. Ser.* **98**, 617–658 (1995).
71. Dye, S. et al. Herschel-ATLAS: modelling the first strong gravitational lenses. *Mon. Not. R. Astron. Soc.* **440**, 2013–2025 (2014).
72. Aravena, M. et al. A survey of the cold molecular gas in gravitationally lensed star-forming galaxies at  $z \geq 2$ . *Mon. Not. R. Astron. Soc.* **457**, 4406–4420 (2016).
73. Venturini, S. & Solomon, P. M. The molecular disk in the Cloverleaf quasar. *Astrophys. J.* **590**, 740–745 (2003).
74. Omont, A. et al.  $\text{H}_2\text{O}$  emission in high- $z$  ultra-luminous infrared galaxies. *Astron. Astrophys.* **551**, A115 (2013).
75. Weiß, A., Henkel, C., Downes, D. & Walter, F. Gas and dust in the Cloverleaf quasar at redshift 2.5. *Astron. Astrophys.* **409**, L41–L45 (2003).
76. Falgarone, E. et al. Large turbulent reservoirs of cold molecular gas around high-redshift starburst galaxies. *Nature* **548**, 430–433 (2017).
77. Vieira, J. D. et al. Dusty starburst galaxies in the early Universe as revealed by gravitational lensing. *Nature* **495**, 344–347 (2013).
78. Ferkinhoff, C. et al. Band-9 ALMA observations of the [N II]  $122\ \mu\text{m}$  line and FIR continuum in two high- $z$  galaxies. *Astrophys. J.* **806**, 260 (2015).
79. Ma, J. et al. Stellar masses and star formation rates of lensed, dusty, star-forming galaxies from the SPT survey. *Astrophys. J.* **812**, 88 (2015).
80. Negrello, M. et al. Herschel-ATLAS: deep HST/WFC3 imaging of strongly lensed submillimetre galaxies. *Mon. Not. R. Astron. Soc.* **440**, 1999–2012 (2014).



**Extended Data Fig. 1 | Velocity-integrated flux maps (moment 0) of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  for SDP 17b.** Black contours show the high-resolution 250-GHz continuum image, obtained from the ALMA archive<sup>76</sup>, with levels of  $3\sigma$ ,  $10\sigma$  and  $50\sigma$  ( $\sigma = 0.6 \times 10^{-1} \text{ mJy beam}^{-1}$ ). Dashed red circles show the adopted apertures for extracting spectra. **a, b**, Images of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  for the  $J=3 \rightarrow 2$  transition. White contours show the

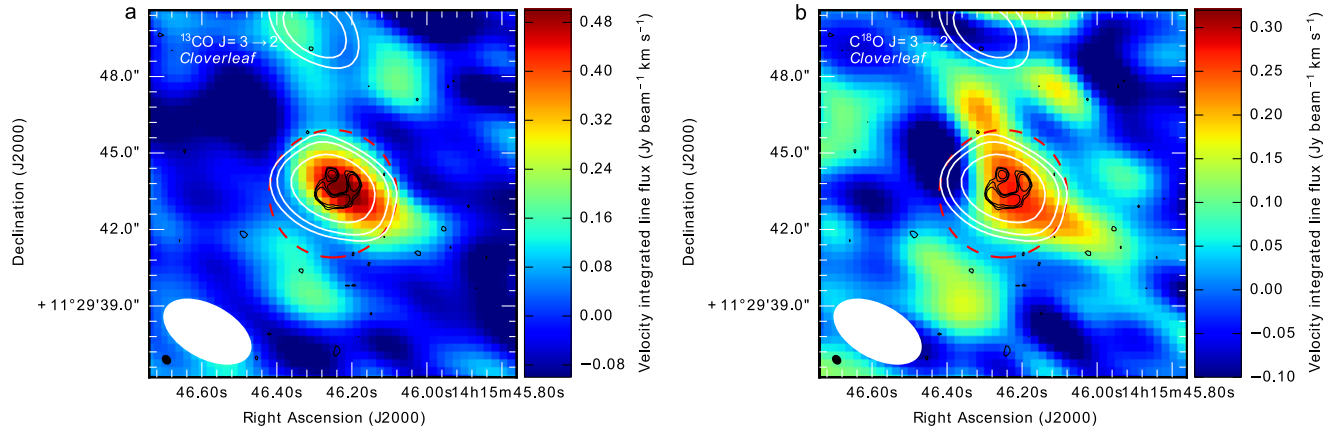
95-GHz continuum, with levels of  $3\sigma$ ,  $5\sigma$  and  $10\sigma$  ( $\sigma = 1.7 \times 10^{-2} \text{ mJy beam}^{-1}$ ). **c, d**, Images of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  for the  $J=4 \rightarrow 3$  transition. White contours show the 133-GHz continuum, with levels of  $3\sigma$ ,  $5\sigma$  and  $10\sigma$  ( $\sigma = 2.3 \times 10^{-2} \text{ mJy beam}^{-1}$ ). The corresponding synthesis beams (white for  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , and black for the 250-GHz continuum) are plotted in the bottom left.



**Extended Data Fig. 2 | Velocity-integrated flux maps (moment 0) of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$   $J=5 \rightarrow 4$  for SPT 0103–45 and the  $J=3 \rightarrow 2$  transition in SPT 0125–47.** Black contours show the high-resolution 336-GHz continuum image, obtained from the ALMA archive<sup>77</sup>, with levels of  $3\sigma$ ,  $10\sigma$  and  $30\sigma$  ( $\sigma = 2.3 \times 10^{-2}$  mJy beam $^{-1}$ ). Dashed red circles show the adopted apertures for extracting spectra. **a, b,** Images of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$   $J=5 \rightarrow 4$  for SPT 0103–45. Blue contours show the narrow  $^{12}\text{CO}$   $J=4 \rightarrow 3$

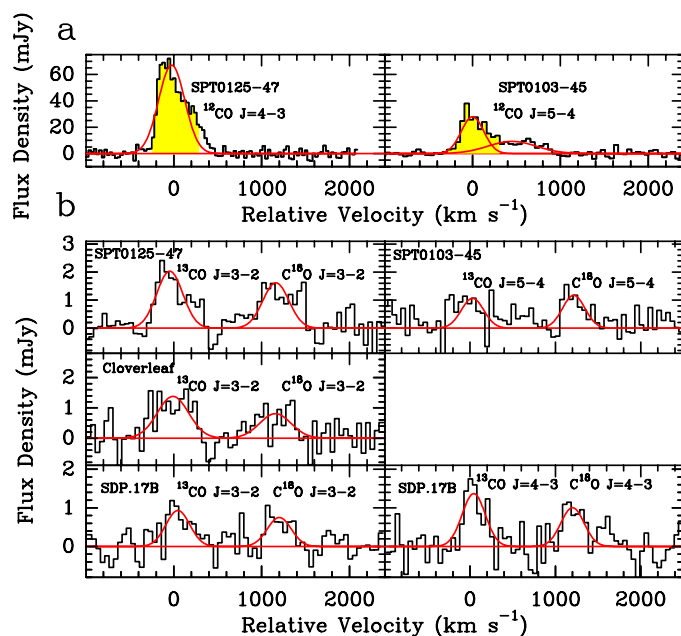
emission, with levels of  $3\sigma$ ,  $10\sigma$  and  $30\sigma$  ( $\sigma = 0.14$  Jy beam $^{-1}$  km s $^{-1}$ ). White contours show the 135-GHz continuum, with levels of  $3\sigma$ ,  $10\sigma$  and  $30\sigma$  ( $\sigma = 2 \times 10^{-2}$  mJy beam $^{-1}$ ). **c, d,** Images of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  for the  $J=3 \rightarrow 2$  transition in SPT 0125–47. White contours show the 94-GHz continuum, with levels of  $3\sigma$ ,  $5\sigma$  and  $10\sigma$  ( $\sigma = 2.2 \times 10^{-2}$  mJy beam $^{-1}$ ). The corresponding synthesis beams (white for  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , and black for the 336-GHz continuum) are plotted in the bottom left.



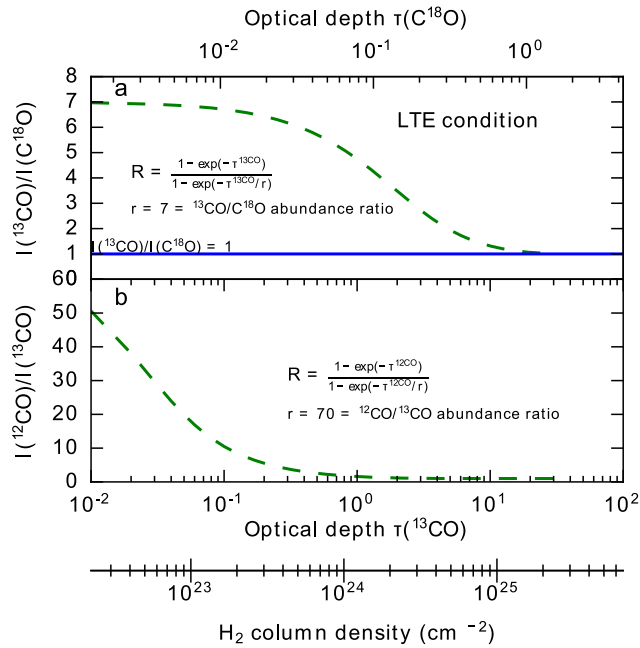


**Extended Data Fig. 3 | Velocity-integrated flux maps (moment 0) of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  for the  $J=3 \rightarrow 2$  transition in the Cloverleaf quasar.** **a**, Image of the  $^{13}\text{CO}$   $J=3 \rightarrow 2$  transition. **b**, Image of the  $\text{C}^{18}\text{O}$   $J=3 \rightarrow 2$  transition. Black contours show the high-resolution 690-GHz continuum image, obtained from the ALMA archive<sup>78</sup>, with levels of  $3\sigma$ ,  $5\sigma$  and  $10\sigma$

( $\sigma = 0.8 \text{ mJy beam}^{-1}$ ). Dashed red circles show the adopted apertures for extracting spectra. White contours show the 92-GHz continuum, with levels of  $3\sigma$ ,  $5\sigma$  and  $10\sigma$  ( $\sigma = 2 \times 10^{-2} \text{ mJy beam}^{-1}$ ). The corresponding synthesis beams (white for  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , and black for the 690-GHz continuum) are plotted in the bottom left.



**Extended Data Fig. 4 | ALMA spectra of the observed  $^{12}\text{CO}$ ,  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  transitions.** **a**, ALMA spectra of  $^{12}\text{CO}$  in SPT 0125-47 and SPT 0103-45. Yellow shading shows the velocity range adopted from  $^{12}\text{CO}$  in the analysis. **b**, ALMA spectra of  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  for all targets. All spectra are in black. Red lines show Gaussian fits to the observed lines. Velocities are labelled relative to their  $^{12}\text{CO}$  or  $^{13}\text{CO}$  transitions.

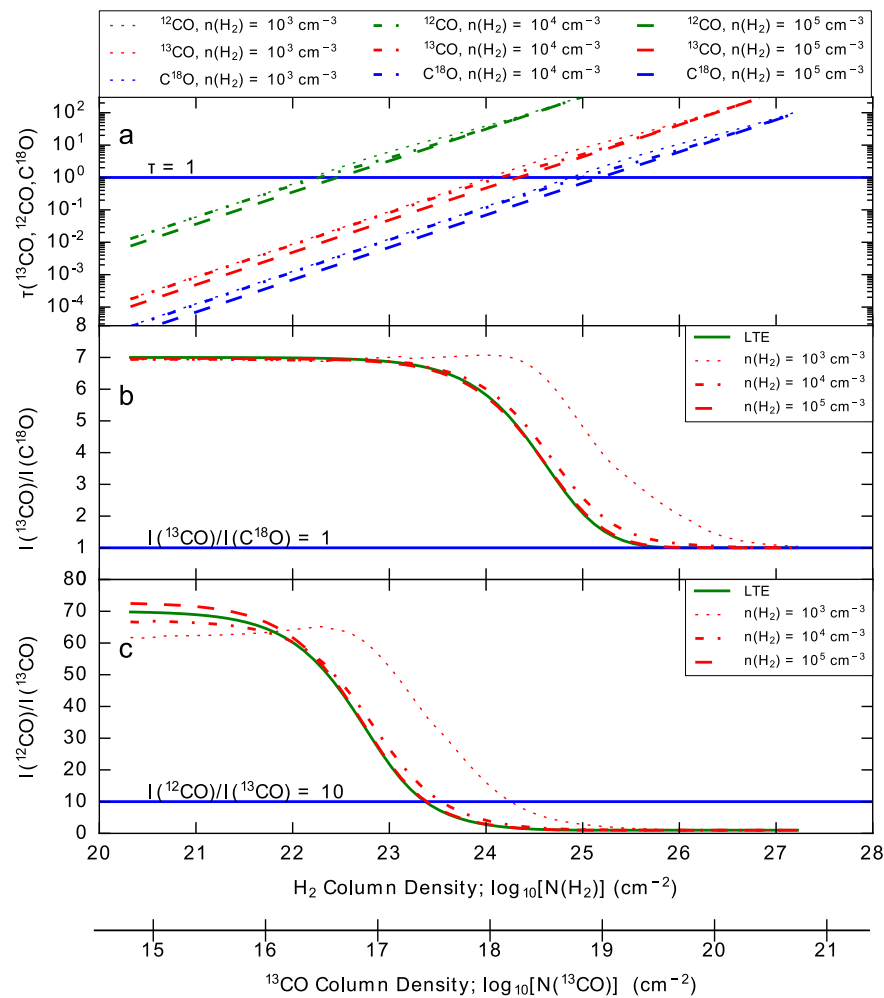


**Extended Data Fig. 5 |  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  and  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratios as a function of optical depth of  $^{13}\text{CO}$ , under LTE conditions.**

**a,**  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  line ratio as a function of optical depth of  $^{13}\text{CO}$ .

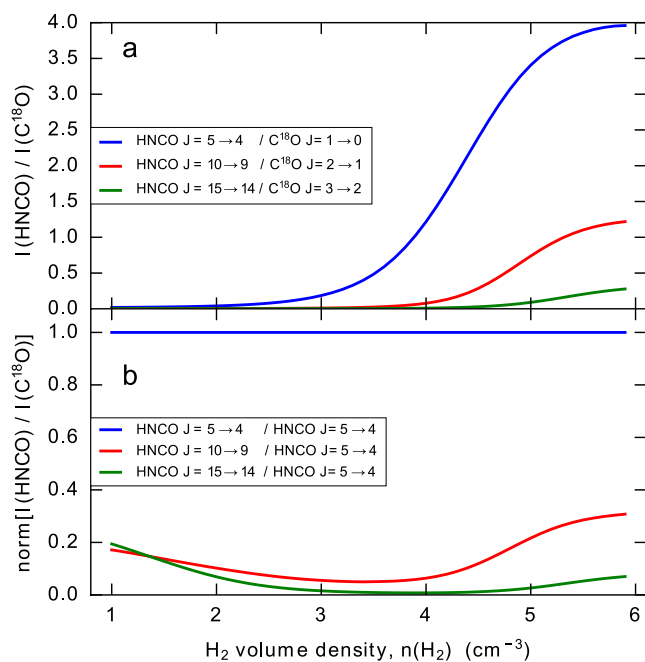
**b,**  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratio as a function of optical depth of  $^{13}\text{CO}$ . Both ratios assume LTE conditions. We assume the abundance ratios of  $^{13}\text{CO}/\text{C}^{18}\text{O}$  and  $^{12}\text{CO}/^{13}\text{CO}$  are 7 and 70, respectively, which are representative values found in the Milky Way. This shows that the  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  line ratio approaches unity (blue line) only when the optical depth of  $\text{C}^{18}\text{O}$  is greater than or equal to 1 (and the corresponding optical depth  $\tau^{13\text{CO}} = 7$ ). The bottom scale bar shows the corresponding  $N_{\text{H}_2}$ , assuming a  $\text{CO}/\text{H}_2$  abundance<sup>78</sup> of  $8.5 \times 10^{-5}$ .  $r$  and  $R$  are the intrinsic abundance ratio and measured line brightness ratio, respectively.





**Extended Data Fig. 6 | Optical depths,  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  and  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratios as a function of  $\text{H}_2$  column density, under non-LTE conditions.** **a**, Optical depths of  $^{12}\text{CO}$ ,  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$ , for the  $J = 3 \rightarrow 2$  transition; **b**,  $I(^{13}\text{CO})/I(\text{C}^{18}\text{O})$  line ratio, and **c**,  $I(^{12}\text{CO})/I(^{13}\text{CO})$  line ratio as a function of  $\text{H}_2$  column density,  $N_{\text{H}_2}$ , and  $^{13}\text{CO}$  column density in various physical conditions, for non-LTE models calculated with RADEX<sup>40</sup>. For all models, we set the abundance ratios of  $^{12}\text{CO}$ ,  $^{13}\text{CO}$  and  $\text{C}^{18}\text{O}$  to be Galactic:  $^{12}\text{CO}/^{13}\text{CO} = 70$  and  $^{13}\text{CO}/\text{C}^{18}\text{O} = 7$ , which are representative values of the Milky Way disk. Different line styles show the

gas conditions of  $\text{H}_2$  volume densities,  $n_{\text{H}_2} = 10^3 \text{ cm}^{-3}$ ,  $10^4 \text{ cm}^{-3}$  and  $10^5 \text{ cm}^{-3}$ . The  $T_{\text{kin}}$  value for all models is set to 30 K, which is a typical dust temperature for the submillimetre galaxy population, and the lowest  $T_{\text{kin}}$  that  $\text{H}_2$  gas can reach for such intensive starburst conditions, due to cosmic ray heating<sup>43</sup>. In **b** and **c**, we also overlay the line ratios (in thick green lines) with the LTE assumption for comparison. All three panels show that for Galactic abundances the line ratio of  $^{13}\text{CO}/\text{C}^{18}\text{O}$  can approach unity only when the  $^{13}\text{CO}$  column density is higher than  $10^{19}$ – $10^{20} \text{ cm}^{-2}$  (that is,  $\text{H}_2$  column density  $N_{\text{H}_2} > 10^{25}$ – $10^{26} \text{ cm}^{-2}$ ).



**Extended Data Fig. 7 |  $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  line ratio and normalized  $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  ratio as a function of  $\text{H}_2$  volume density.**

**a,**  $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  line ratio as a function of  $\text{H}_2$  volume density.

**b,**  $I(\text{HNCO})/I(\text{C}^{18}\text{O})$  line ratio as a function of  $\text{H}_2$  volume density, normalized with  $I(\text{HNCO } J = 5 \rightarrow 4)/I(\text{C}^{18}\text{O } J = 1 \rightarrow 0)$ . Both ratios are calculated using RADEX<sup>40</sup>, in which we assume the same abundances as measured in Arp 220<sup>47</sup>. We assume  $T_{\text{kin}} = 30 \text{ K}$  as the representative kinetic temperature of the  $\text{H}_2$  gas.

**Extended Data Table 1 | Target properties**

Short name	IAU name	R.A. J2000	Dec. J2000	Redshift $z$	Lensing amplification, $\mu$	$L_{\text{IR}}/\mu$ $10^{13} L_{\odot}$	$M_{\star}/\mu$ $10^{10} M_{\odot}$
SPT 0103–45	SPT-S J010312–4538.8	01:03:11.50	–45:38:53.9	3.0917	$5.3 \pm 0.11$	1.2	$5.5^{+6.1}_{-2.9}$
SPT 0125–47	SPT-S J012506–4723.7	01:25:07.08	–47:23:56.0	2.5148	$5.5 \pm 0.1^{\star}$	2.2	–
SDP.17b	HATLAS J090302.9014127	09:03:03.02	–01:41:26.9	2.3051	$3.56^{+0.19}_{-0.17}$	2.0	$24.2^{+8.6}_{-4.0}$
Cloverleaf	H1413+117	14:15:46.23	+11:29:44.0	2.5585	$11^{\dagger}$	6.0	–

Data for SDP.17b are from refs <sup>71,80</sup>. IAU, International Astronomical Union, R.A., right ascension, Dec., declination.

<sup>\*</sup>We adopt an amplification factor,  $\mu$ , derived from lens modelling using 850- $\mu\text{m}$  ALMA visibility data in the literature<sup>72,79</sup>.

<sup>†</sup>We adopt the best-determined amplification factor,  $\mu$ , reported in the literature<sup>35,73</sup>, whose lensing model is derived from CO  $J=7 \rightarrow 6$  line emission, and could better reproduce a single source on the lens plane<sup>73</sup>. Unfortunately the uncertainty of this amplification factor was not reported, but the uncertainty in  $\mu$  does not jeopardise our conclusions.



Extended Data Table 2 | ALMA observational information

Target	SPT 0103–45	SPT 0103–45	SPT 0125–47	SPT 0125–47	SDP.17B	SDP.17B	Cloverleaf
Isotopologue	$^{13}\text{CO}$ , $\text{C}^{18}\text{O}$	$^{12}\text{CO}$	$^{13}\text{CO}$ , $\text{C}^{18}\text{O}$	$^{12}\text{CO}$	$^{13}\text{CO}$ , $\text{C}^{18}\text{O}$	$^{13}\text{CO}$ , $\text{C}^{18}\text{O}$	$^{13}\text{CO}$ , $\text{C}^{18}\text{O}$
Transition	$J = 5 \rightarrow 4$	$J = 5 \rightarrow 4$	$J = 3 \rightarrow 2$	$J = 4 \rightarrow 3$	$J = 3 \rightarrow 2$	$J = 4 \rightarrow 3$	$J = 3 \rightarrow 2$
Observing Date	21-Jan-2016	21-Jan-2016	21-Jan-2016	21-Jan-2016	17-Jan-2016	16-Jan-2016	08-Apr-2016
Bandpass Calibrator	J2357–5311	J2357–5311	J2357–5311	J2357–5311	J0854+2006	J0854+2006	J1337–1257
Flux Calibrator	Neptune	J2357–5311	Neptune	Neptune	J0854+2006	J0854+2006	Callisto
Gain Calibrator	J0056–4451	J0051–4226	J0124–5113	J0124–5113	J0909+0121	J0909+0121	J1415+1320
Integration Time (s)	1300	120	605	120	816	726	1753
Median PWV (mm)	6.1	5.6	6.2	6.0	2.0	3.2	3.0
Median $T_{\text{sys}}$ (K)	86	85	93	88	52	60	70
Angular Resolution	$2.5'' \times 1.8''$	$2.7'' \times 1.6''$	$3.6'' \times 2.5''$	$2.5'' \times 1.8''$	$3.1'' \times 2.3''$	$2.7'' \times 1.8''$	$3.7'' \times 2.0''$

PWV, precipitable water vapour.

Extended Data Table 3 | Observed targets, lines, frequencies, linewidths and fluxes

Target	Transition $J \rightarrow J-1$	$\nu_{\text{obs}}$ GHz	$I_{\text{line}}$ Jy km s <sup>-1</sup>	$I_{\text{line}}^{\text{mom0}}$ Jy km s <sup>-1</sup>	$\sigma^{\text{theo}} \star$ Jy km s <sup>-1</sup>	$\Delta V_{\text{line}}$ km s <sup>-1</sup>	$F_{\text{peak}}$ mJy
SDP.17b	<sup>12</sup> CO $J = 4 \rightarrow 3$	139.49	9.1 ± 0.3	—	—	320	~40
SDP.17b	<sup>13</sup> CO $J = 3 \rightarrow 2$	100.02	0.32 ± 0.05	0.34 ± 0.08	0.08		0.9 ± 0.3
SDP.17b	C <sup>18</sup> O $J = 3 \rightarrow 2$	99.64	0.26 ± 0.05	0.32 ± 0.08	0.08		0.8 ± 0.3
SDP.17b	<sup>13</sup> CO $J = 4 \rightarrow 3$	133.36	0.47 ± 0.07	0.46 ± 0.08	0.07		1.3 ± 0.4
SDP.17b	C <sup>18</sup> O $J = 4 \rightarrow 3$	132.85	0.34 ± 0.06	0.50 ± 0.08	0.07		1.0 ± 0.4
Cloverleaf	<sup>12</sup> CO $J = 3 \rightarrow 2$	97.17	13.2 ± 0.2	—	—	400	30 ± 1.7
Cloverleaf	<sup>13</sup> CO $J = 3 \rightarrow 2$	92.90	0.65 ± 0.09	0.61 ± 0.06	0.07		1.4 ± 0.4
Cloverleaf	C <sup>18</sup> O $J = 3 \rightarrow 2$	92.55	0.40 ± 0.10	0.43 ± 0.06	0.07		0.8 ± 0.4
SPT 0103–45	<sup>12</sup> CO $J = 4 \rightarrow 3$	112.68	8.2 ± 0.6	—	—		32 ± 0.6
SPT 0103–45	<sup>12</sup> CO $J = 5 \rightarrow 4$	140.91	8.8 ± 0.5 <sup>†</sup>	8.8 ± 0.6 <sup>†</sup>	—	300 <sup>†</sup>	27.8 ± 0.2
SPT 0103–45	<sup>13</sup> CO $J = 5 \rightarrow 4$	134.65	0.37 ± 0.07	0.38 ± 0.05	0.07		1.2 ± 0.4
SPT 0103–45	C <sup>18</sup> O $J = 5 \rightarrow 4$	134.13	0.35 ± 0.09	0.39 ± 0.07	0.07		1.2 ± 0.4
SPT 0125–47	<sup>12</sup> CO $J = 3 \rightarrow 2$	98.38	18.1 ± 0.5	18.0 ± 0.5	—		43 ± 4
SPT 0125–47	<sup>12</sup> CO $J = 4 \rightarrow 3$	131.21	26.9 ± 0.7	26.8 ± 0.7	—	400	69 ± 3
SPT 0125–47	<sup>13</sup> CO $J = 3 \rightarrow 2$	94.06	0.78 ± 0.09	0.86 ± 0.07	0.1		2.0 ± 0.4
SPT 0125–47	C <sup>18</sup> O $J = 3 \rightarrow 2$	93.70	0.63 ± 0.07	0.71 ± 0.1	0.1		1.6 ± 0.4

Data are from refs <sup>32,74,75</sup>.  $T_{\text{sys}}$  is the system temperature.

\*Theoretical noise level calculated using the ALMA sensitivity calculator (<https://almascience.eso.org/proposing/sensitivity-calculator>).

<sup>†</sup>There are two velocity components in the <sup>12</sup>CO spectrum. We adopt only the narrow component, seen for <sup>12</sup>CO  $J = 5 \rightarrow 4$ , to avoid the broad and weaker component (see Extended Data Fig. 4).

# Deterministic quantum state transfer and remote entanglement using microwave photons

P. Kurpiers<sup>1,4\*</sup>, P. Magnard<sup>1,4</sup>, T. Walter<sup>1</sup>, B. Royer<sup>2</sup>, M. Pechal<sup>1</sup>, J. Heinsoo<sup>1</sup>, Y. Salathé<sup>1</sup>, A. Akin<sup>1</sup>, S. Storz<sup>1</sup>, J.-C. Besse<sup>1</sup>, S. Gasparinetti<sup>1</sup>, A. Blais<sup>2,3</sup> & A. Wallraff<sup>1\*</sup>

**Sharing information coherently between nodes of a quantum network is fundamental to distributed quantum information processing. In this scheme, the computation is divided into subroutines and performed on several smaller quantum registers that are connected by classical and quantum channels<sup>1</sup>. A direct quantum channel, which connects nodes deterministically rather than probabilistically, achieves larger entanglement rates between nodes and is advantageous for distributed fault-tolerant quantum computation<sup>2</sup>. Here we implement deterministic state-transfer and entanglement protocols between two superconducting qubits fabricated on separate chips. Superconducting circuits<sup>3</sup> constitute a universal quantum node<sup>4</sup> that is capable of sending, receiving, storing and processing quantum information<sup>5–8</sup>. Our implementation is based on an all-microwave cavity-assisted Raman process<sup>9</sup>, which entangles or transfers the qubit state of a transmon-type artificial atom<sup>10</sup> with a time-symmetric itinerant single photon. We transfer qubit states by absorbing these itinerant photons at the receiving node, with a probability of  $98.1 \pm 0.1$  per cent, achieving a transfer-process fidelity of  $80.02 \pm 0.07$  per cent for a protocol duration of only 180 nanoseconds. We also prepare remote entanglement on demand with a fidelity as high as  $78.9 \pm 0.1$  per cent at a rate of 50 kilohertz. Our results are in excellent agreement with numerical simulations based on a master-equation description of the system. This deterministic protocol has the potential to be used for quantum computing distributed across different nodes of a cryogenic network.**

Remote entanglement has been realized probabilistically using heralded or unheralded protocols based on measurement projection<sup>11–14</sup>, single-<sup>15,16</sup> or two-photon<sup>17–20</sup> detection or direct transfer of a single photon<sup>21,22</sup>. See Methods and Extended Data Fig. 1 for an overview of selected experimental results, including a discussion of concurrent deterministic experiments performed with superconducting circuits<sup>23,24</sup>. However, a fully deterministic implementation<sup>25</sup> of direct transfer protocols is more challenging to realize. In the protocol<sup>25</sup>, a stationary atom is coupled to a single-mode cavity in remote quantum nodes and a coherent drive entangles the state of the atom with the field of the cavity. The cavity is coupled to a directional quantum channel into which the field is emitted as a time-symmetric single photon. This photon travels to the receiving node where it is ideally absorbed with unit probability using the time-reversed coherent drive (Fig. 1a). In addition to establishing entanglement between the nodes, direct transfer of quantum information offers the possibility to transmit arbitrary qubit states from one node to the other.

In our adaptation of this scheme (Fig. 1b) to the circuit quantum electrodynamic architecture, each quantum node (labelled A and B) is composed of a superconducting transmon qubit with transition frequency  $\nu_{\text{ge}}^{\text{A}} = 6.343$  GHz or  $\nu_{\text{ge}}^{\text{B}} = 6.093$  GHz dispersively coupled to two coplanar microwave resonators, analogous to an atom coupled to two cavity modes. One resonator is dedicated to dispersive transmon readout and the other to excitation transfer. The transfer resonators at the

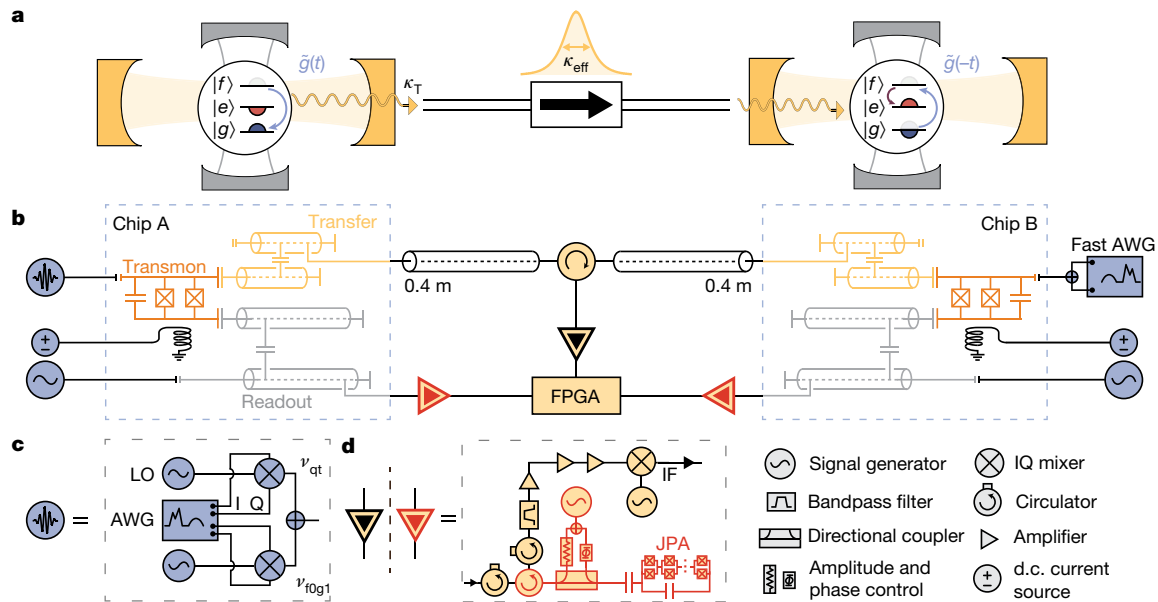
two nodes are tuned to have matching frequencies  $\nu_{\text{T}} \approx 8.400$  GHz and large bandwidths  $\kappa_{\text{T}}/(2\pi)$  of the order of 10 MHz (see Methods). All resonators are coupled to dedicated filters, to protect the transmons from Purcell decay<sup>26,27</sup>. An external coaxial line with a length of 0.9 m, bisected with a circulator, connects the transfer circuits of both chips. With this set-up, photons are routed from node A to B and from node B to a detection line. If perfect absorption of the photon can be realized and independent detection of the photon is not needed or desired, then the circulator can be omitted from the circuit<sup>25</sup>. To generate a controllable light–matter interaction, we apply a coherent microwave tone to the transmon, which induces an effective interaction  $\tilde{g}(t)$  with tunable amplitude and phase<sup>9,28</sup> between states  $|f, 0\rangle$  and  $|g, 1\rangle$ . Here,  $|s, n\rangle$  denotes a Jaynes–Cummings dressed eigenstate with transmon state  $|s\rangle$  and Fock state of the transfer resonator  $|n\rangle$ . The two lowest-energy eigenstates ( $|g\rangle$  and  $|e\rangle$ ) of the transmon form the qubit subspace; the second excited state ( $|f\rangle$ ) is used as an auxiliary level to control the light–matter interaction in our experiment. This interaction swaps an excitation from the transmon to the transfer resonator, which then couples to a mode propagating towards node B. By controlling  $\tilde{g}(t)$  (see Methods), we shape the itinerant photon to have a time-symmetric envelope  $\phi(t) = \frac{1}{2} \sqrt{\kappa_{\text{eff}}} \text{sech}(\kappa_{\text{eff}} t / 2)$ , with an adjustable photon bandwidth  $\kappa_{\text{eff}}$  limited only by  $\kappa_{\text{T}}$ . By inducing the reverse process  $|g, 1\rangle \leftrightarrow |f, 0\rangle$  with the time-reversed amplitude and phase profile of  $\tilde{g}(t)$ , we absorb the itinerant photon in the transmon at node B. Ideally, this procedure returns all photonic modes to their vacuum state. We note that in our system this process could also be implemented with asymmetric photon shapes, or ones with a more structured time dependence<sup>29</sup>, as long as the physical constraints on the bandwidth required for its emission and absorption are met at the respective sites.

To characterize the excitation transfer, we start by initializing the transmon in its ground state<sup>30</sup>, after which we apply a sequence of two  $\pi$  pulses ( $R_{\text{ge}}^{\pi}, R_{\text{ef}}^{\pi}$ ) to prepare the transmon at the receiving node B in state  $|f, 0\rangle$ . Next, we induce the effective coupling  $\tilde{g}(t)$  with a modulated drive  $R_{\text{f0g1}}^{\tau}$  to emit a symmetric photon<sup>9</sup> (Fig. 2a). We vary the instantaneous frequency of  $R_{\text{f0g1}}^{\tau}$  to compensate for the drive-amplitude-dependent a.c. Stark shift of the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition (see Methods). Here, and in all subsequent measurements, the population of the transmon states are extracted using single-shot readout with a correction to account for measurement errors (see Methods). The populations of the three lowest levels of the transmon  $P_{\text{g,e,f}}$  are measured immediately after truncating the emission pulse  $R_{\text{f0g1}}^{\tau}$  at time  $\tau$  (Fig. 2b). In this way, we observe that the transmon evolves smoothly from  $|f\rangle$  to  $|g\rangle$  during the emission process. At the end of the protocol, the emitting transmon reaches a ground-state population of  $P_{\text{g}} = 95.8\%$ , which characterizes the emission efficiency.

To verify that the envelope of the emitted photon has the target shape and bandwidth  $\kappa_{\text{eff}}^{\text{B}}/(2\pi) = 10.6$  MHz, we repeat the emission protocol with an initial transmon state  $(|g\rangle + |f\rangle)/\sqrt{2}$  and measure the averaged electric-field amplitude  $\langle a_{\text{out}}(t) \rangle \propto \phi(t)$  of the emitted photon

<sup>1</sup>Department of Physics, ETH Zürich, Zürich, Switzerland. <sup>2</sup>Institut Quantique and Département de Physique, Université de Sherbrooke, Sherbrooke, Québec, Canada. <sup>3</sup>Canadian Institute for Advanced Research, Toronto, Ontario, Canada. <sup>4</sup>These authors contributed equally: P. Kurpiers, P. Magnard. \*e-mail: philipp.kurpiers@phys.ethz.ch; andreas.wallraff@phys.ethz.ch





**Fig. 1 | Schematic and measurement set-up. a**, Quantum optical schematic of a deterministic unidirectional entanglement protocol between two cavity quantum electrodynamic nodes of a quantum network. At the first node, a three-level system is prepared in its second excited state  $|f\rangle$  (grey half-circle) and driven coherently ( $\tilde{g}(t)$ , blue arrow) to  $|g\rangle$  (blue half-circle), creating the transfer cavity field  $|1\rangle$  (light yellow). The cavity field couples into the directional quantum channel with rate  $\kappa_T$  as a single-photon wavepacket with an effective bandwidth  $\kappa_{\text{eff}}$  (yellow hyperbolic secant shape). In the second quantum node, the time-reversed drive  $\tilde{g}(-t)$  transfers the excitation from  $|g\rangle$  to  $|f\rangle$  in the presence of the transferred photon field  $|1\rangle$ . Finally, the protocol is completed with a transfer pulse between  $|f\rangle$  and  $|e\rangle$  (red half-circle) to return to the qubit subspace. In addition, each three-level system is coupled to a readout cavity (grey). **b**, Implementation of the system depicted in **a** in a planar, chip-based, circuit quantum electrodynamic architecture (Extended Data Fig. 2). At each node, a transmon (orange) is coupled capacitively to two  $\lambda/4$  coplanar waveguide resonators and Purcell filter circuits<sup>27</sup> that act as the transfer (yellow) and readout (grey) cavities, respectively. The output transmission lines are coupled galvanically to the corresponding circuit. A directional

quantum channel is realized using a semi-rigid coaxial cable and a circulator connecting to the output port of the transfer circuit Purcell filter at each node. **c**, **d**, Details of the circuit quantum electrodynamic implementation. **c**, Combined qutrit ( $\nu_{qt}$ ) and  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition ( $\nu_{fg1}$ ) microwave drive using single-side-band modulation with in-phase (I) and quadrature (Q) mixers driven by a local oscillator (LO) and with an envelope defined by an arbitrary-waveform generator (AWG) for node A. At node B, these drives are synthesized directly by a fast AWG with  $25 \text{ GS s}^{-1}$ . **d**, Schematic of microwave detection lines (black, red triangles). All detection lines consist of two isolators, a bandpass filter, a cryogenic amplifier (HEMT) and two room-temperature amplifiers followed by a filter and analogue down-conversion to an intermediate frequency of 250 MHz. The down-converted signal is lowpass-filtered, digitized using an analogue-to-digital converter and recorded using a field-programmable gate array (FPGA). The transmon-readout lines include an additional Josephson parametric amplifier (JPA) circuit (red elements) between the first two isolators. The JPA is pumped by a signal generator and the reflected pump signal from the JPA is cancelled at a directional coupler using amplitude- and phase-controlled destructive interference.

state  $(|0\rangle + |1\rangle)/\sqrt{2}$  using heterodyne detection<sup>31</sup> (Fig. 2c). We prepare this photon state because of its non-zero average electric field<sup>9</sup>.

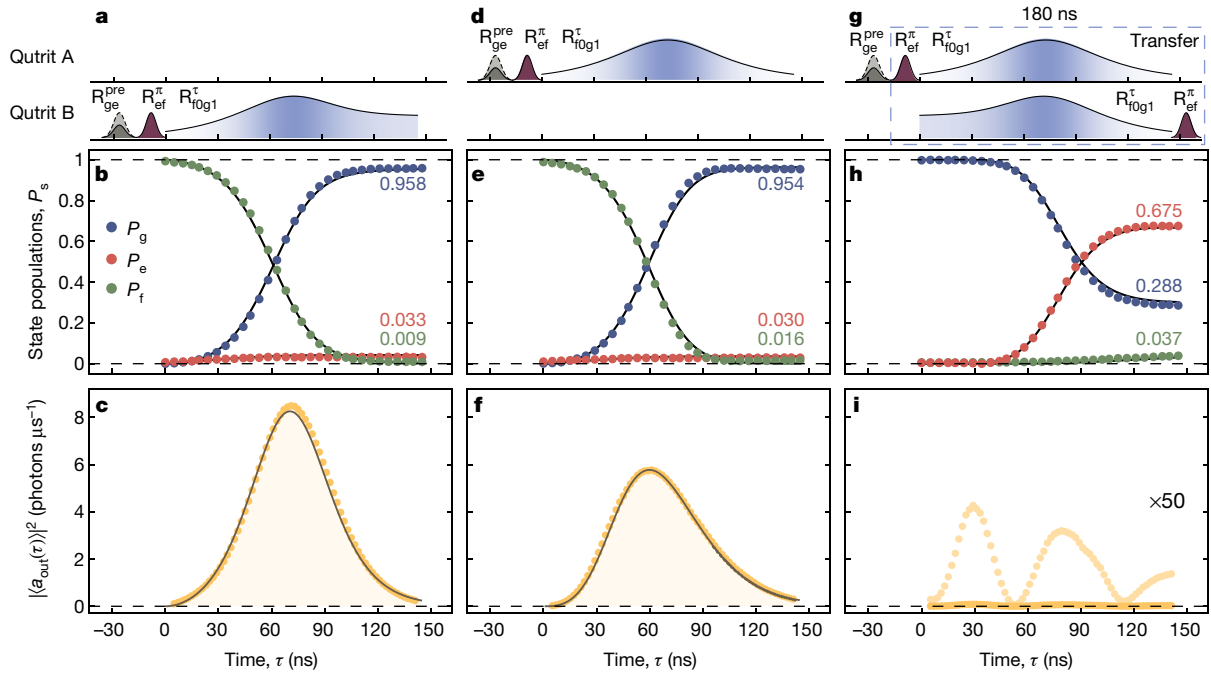
Repeating the emission protocol from node A leads to similar dynamics of the transmon population (Fig. 2e). We adjust the amplitude and phase of the transfer pulse (Fig. 2d) so that the photons emitted from each node A and B have similar effective bandwidth  $\kappa_{\text{eff}}$  in spite of their respective transfer resonator bandwidths  $\kappa_T$  differing by approximately 30% (see Methods). The detected integrated power  $\int |\langle a_{\text{out}}(t) \rangle|^2 dt$  of the photon emitted from node A (Fig. 2f) is  $I_{AB} = 23.0\% \pm 0.5\%$  lower than that emitted from node B owing to loss accumulated as the photon travels from node A to B. The photon loss  $I_{AB}$  is extracted from the ratio of the integrated photon powers for emission from nodes B and A (see Methods). In addition, the envelope of the photon emitted from node A is slightly distorted by the reflection off node B, as determined by the response function of its transfer resonator, which is fully captured by our theoretical model.

To characterize the absorption of the single time-symmetric photon emitted from node A at the receiving node by time-reversing the emission pulse of node B (Fig. 2a, g), we measure the population of transmon B during the process. We apply a  $\pi$  pulse to transmon B to map  $|f\rangle$  back to the qubit subspace before performing the readout. We observe the population of  $|e\rangle$  to rise smoothly and saturate at  $P_e^{\text{sat}} = 67.5\%$  (Fig. 2h). This saturation level reflects the efficiency of the protocol for the transfer of a single excitation (a single photon), which is executed in a pulse sequence of only 180-ns duration (Fig. 2g). From the ratio of the integrated power of the emitted photon in the absence

(Fig. 2i) or presence (Fig. 2f) of the absorption pulse, the absorption efficiency is determined to reach  $98.1\% \pm 0.1\%$ .

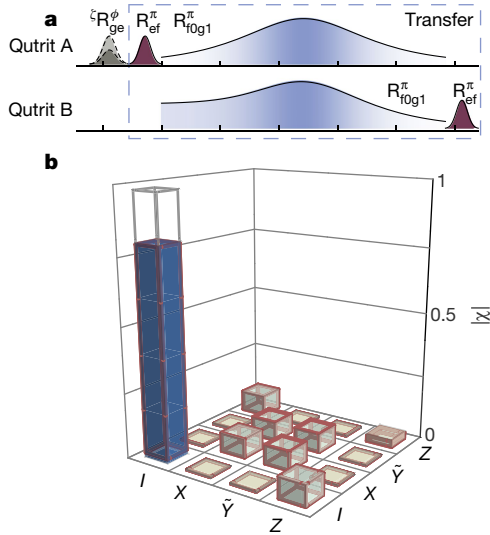
The results of master-equation simulations of the excitation transfer (solid lines in Fig. 2), using parameters extracted from independent measurements (see Methods), display excellent agreement with the measured data. This demonstrates a high level of control over the emission and absorption processes and an accurate understanding of the experimental imperfections dominated by qutrit decoherence and photon loss.

We demonstrate the use of our protocol to transfer deterministically an arbitrary qubit state from node A over a distance of about 0.9 m along a coaxial line to node B. This is realized by preparing the receiving transmon (B) in state  $|g\rangle$ , applying a  $R_{\pi}^x$  pulse to the sending transmon (A), followed by the emission or absorption pulse and finally a rotation  $R_{\pi}^x$  on transmon B. We characterize the quantum state transfer by reconstructing its process matrix  $\chi$  with quantum process tomography (Fig. 3b). For that purpose, we prepare all six mutually unbiased qubit basis states<sup>32</sup> at node A, transfer them to node B, and reconstruct the transferred state using quantum state tomography (see Methods). We determine a process fidelity of  $\mathcal{F}_p = \text{tr}(\chi \chi_{\text{ideal}}) = 80.02\% \pm 0.07\%$ , well above the limit of  $1/2$  that can be achieved using local gates and classical communication only. The process matrix  $\chi_{\text{sim}}$  calculated with the master-equation simulations agrees very well with the data (absolute values shown as red outlines in Fig. 3b). This is supported by the small trace distance<sup>33</sup>  $\text{tr}|\chi - \chi_{\text{sim}}|/2 = 0.015$ , which ideally is 0 for identical process matrices and 1 for orthogonal ones.



**Fig. 2 | Emission, transfer and absorption of a single photon.** **a, d,** The transmons at node B (**a**) and node A (**d**) are prepared in the state  $|f\rangle$  using Gaussian derivative removal by adiabatic gate (DRAG) microwave pulses  $R_{ge}^{pre=\pi}$  and  $R_{ef}^{\pi}$ . **b, e,** We characterize (filled circles) the time dependence ( $\tau$ ) of the qutrit populations  $P_{g,e,f}$  while driving the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition. The phase (white–blue shading) of the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition drive is modulated to compensate the drive-induced quadratic a.c. Stark shift. **c, f,** The mean field amplitude squared  $|\langle a_{out}(\tau) \rangle|^2$  of the travelling photons emitted from node B (**c**) and node A (**f**) is obtained for the photon

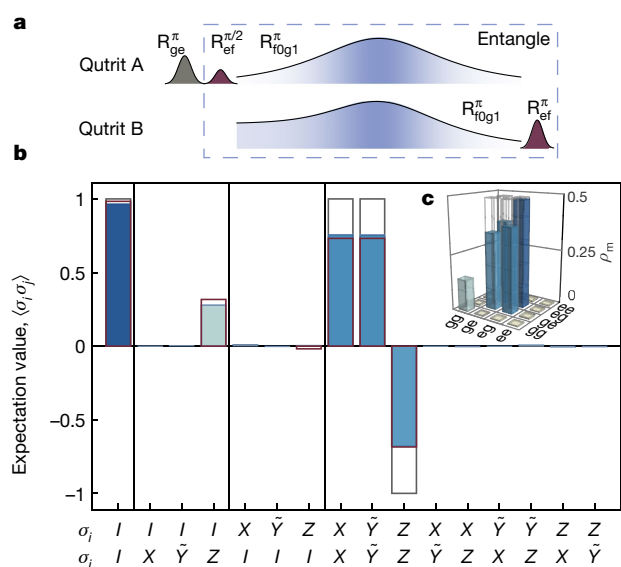
state  $(|0\rangle + |1\rangle)/\sqrt{2}$  that is emitted by preparing each transmon in  $(|g\rangle + |f\rangle)/\sqrt{2}$  ( $R_{ge}^{pre=\pi/2}$  in **a** and **d**). The effective photon bandwidths are adjusted to be  $\kappa_{eff}^A/(2\pi) = 10.4$  MHz and  $\kappa_{eff}^B/(2\pi) = 10.6$  MHz. The solid lines in **b, c, e, f, h** and **i** are results of master-equation simulations (see text for details). **g,** Excitation transfer protocol. **h,** The time dependence of  $P_{g,e,f}$  when executing the excitation transfer protocol from qubit A to qubit B with  $R_{ge}^{pre=\pi}$ . **i,** The residual  $|\langle a_{out}(\tau) \rangle|^2$  (light yellow, multiplied by 50) during the absorption process ( $R_{ge}^{pre=\pi/2}$  in **g**).



**Fig. 3 | Quantum state transfer.** **a,** Pulse scheme used to characterize the qubit state transfer between the two nodes. We prepare six mutually unbiased input states with rotations  ${}^xR_{ge}^0$ ,  ${}^xR_{ge}^{\pi/2}$ ,  ${}^xR_{ge}^{-\pi/2}$ ,  ${}^yR_{ge}^{\pi/2}$ ,  ${}^yR_{ge}^{-\pi/2}$  and  ${}^zR_{ge}^{\pi}$  at node A (denoted by  ${}^{\zeta}R_{ge}^{\phi}$  where  $\zeta$  is the rotation axis). **b,** We experimentally obtain a process matrix (absolute value  $|\chi|$  shown as coloured bars) in the basis of the Pauli matrices  $I, X = \sigma_x, Y = i\sigma_y$ , and  $Z = \sigma_z$  with a fidelity of  $\mathcal{F}_p = 80.02\% \pm 0.07\%$  relative to the ideal identity operation. The grey and red outlines show the ideal value and the master-equation simulation of the absolute values of the process matrix, respectively. The trace distance between the measurement and the simulation is 0.015.

Furthermore, we use the excitation transfer to generate two-qubit remote-entangled states between nodes A and B deterministically. The protocol starts by preparing transmons A and B in states  $(|e\rangle + |f\rangle)/\sqrt{2}$  and  $|g\rangle$ , respectively, and then applying the emission or absorption pulses followed by a rotation  $R_{ef}^{\pi}$  on transmon B to generate the entangled Bell state  $|\psi^+\rangle = (|e, g\rangle + |g, e\rangle)/\sqrt{2}$ . Because leakage to the  $|f\rangle$  level at both nodes leads to errors in the two-qubit density matrix reconstruction, we extract the full two-qutrit density matrix  $\rho_{3\otimes 3}$  from quantum state tomography experiments (see Methods). For illustration purposes, we display the two-qubit density matrix  $\rho_m$  (Fig. 4b, c), which consists of the two-qubit elements of  $\rho_{3\otimes 3}$ . We find a state fidelity of  $\mathcal{F}_{|\psi^+\rangle}^s = \langle \psi^+ | \rho_m | \psi^+ \rangle = 78.9\% \pm 0.1\%$  compared to the ideal Bell state, and a concurrence  $\mathcal{C}(\rho_m) = 0.747 \pm 0.004$  (see Methods for a detailed discussion). The density matrix  $\rho_{sim}$  calculated from the master-equation simulations of the entanglement protocol (red outlines in Fig. 4) is in excellent agreement with the experimental results, displaying a small trace distance of  $\text{tr}|\rho_m - \rho_{sim}|/2 = 0.024$ . We decompose the infidelity into approximately 10.5% from photon loss, 9% from finite transmon coherence times and 2% from pulse truncation.

Using transmons with relaxation and coherence times of  $T_{1ge} = T_{2ge} = 30$   $\mu\text{s}$  and  $T_{1ef} = T_{2ef} = 20$   $\mu\text{s}$ , and with an achievable 12% loss between the nodes, we calculate that our protocol would enable deterministic generation of remote-entangled states with a fidelity of 93%. In addition, we expect our protocol to be extendable to quantum network applications to generate deterministic heralded remote entanglement<sup>4</sup>, using the three-level structure of the transmons and encoding quantum information in different time bins to detect photon loss. These perspectives indicate that the approach demonstrated here may serve as



**Fig. 4 | Remote-entanglement generation.** **a**, Pulse scheme to generate deterministic remote entanglement between nodes A and B. **b**, Expectation values of two-qubit Pauli operators  $\langle \sigma_i \sigma_j \rangle$ . The coloured bars indicate the measurement results; the ideal expectation values for the Bell state  $|\psi^+\rangle = (|e, g\rangle + |g, e\rangle)/\sqrt{2}$  and the results of a master-equation simulation are shown as grey and red outlines, respectively. We calculate a fidelity of  $\mathcal{F}^s_{|\psi^+\rangle} = 78.9\% \pm 0.1\%$ , which is well explained by photon loss and decoherence. **c**, Reconstructed density matrix  $\rho_m$  after execution of the remote-entanglement protocol.

the basis for distributed quantum computation in the circuit quantum electrodynamic architecture using distinct cryogenic nodes.

### Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0195-y>.

Received: 29 December 2017; Accepted: 27 March 2018;  
Published online 13 June 2018.

1. Cirac, J. I., Ekert, A. K., Huelga, S. F. & Macchiavello, C. Distributed quantum computation over noisy channels. *Phys. Rev. A* **59**, 4249–4254 (1999).
2. Jiang, L., Taylor, J. M., Sørensen, A. S. & Lukin, M. D. Distributed quantum computation based on small quantum registers. *Phys. Rev. A* **76**, 062323 (2007).
3. Wallraff, A. et al. Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431**, 162–167 (2004).
4. Reiserer, A. & Rempe, G. Cavity-based quantum networks with single atoms and optical photons. *Rev. Mod. Phys.* **87**, 1379–1418 (2015).
5. Eichler, C. et al. Observation of entanglement between itinerant microwave photons and a superconducting qubit. *Phys. Rev. Lett.* **109**, 240501–240505 (2012).
6. Wenner, J. et al. Catching time-reversed microwave coherent state photons with 99.4% absorption efficiency. *Phys. Rev. Lett.* **112**, 210501 (2014).
7. Johnson, B. R. et al. Quantum non-demolition detection of single microwave photons in a circuit. *Nat. Phys.* **6**, 663–667 (2010).
8. DiCarlo, L. et al. Demonstration of two-qubit algorithms with a superconducting quantum processor. *Nature* **460**, 240–244 (2009).
9. Pechal, M. et al. Microwave-controlled generation of shaped single photons in circuit quantum electrodynamics. *Phys. Rev. X* **4**, 041010 (2014).
10. Koch, J. et al. Charge-insensitive qubit design derived from the Cooper pair box. *Phys. Rev. A* **76**, 042319 (2007).
11. Julsgaard, B., Kozhekin, A. & Polzik, E. S. Experimental long-lived entanglement of two macroscopic objects. *Nature* **413**, 400–403 (2001).

12. Chou, C. W. et al. Measurement-induced entanglement for excitation stored in remote atomic ensembles. *Nature* **438**, 828–832 (2005).
13. Choi, K. S., Goban, A., Papp, S. B., van Enk, S. J. & Kimble, H. J. Entanglement of spin waves among four quantum memories. *Nature* **468**, 412–416 (2010).
14. Roch, N. et al. Observation of measurement-induced entanglement and quantum trajectories of remote superconducting qubits. *Phys. Rev. Lett.* **112**, 170501 (2014).
15. Slodička, L. et al. Atom-atom entanglement by single-photon detection. *Phys. Rev. Lett.* **110**, 083603 (2013).
16. Delteil, A. et al. Generation of heralded entanglement between distant hole spins. *Nat. Phys.* **12**, 218–223 (2016).
17. Moehring, D. L. et al. Entanglement of single-atom quantum bits at a distance. *Nature* **449**, 68–71 (2007).
18. Lee, K. C. et al. Entangling macroscopic diamonds at room temperature. *Science* **334**, 1253–1256 (2011).
19. Hofmann, J. et al. Heralded entanglement between widely separated atoms. *Science* **337**, 72–75 (2012).
20. Bernien, H. et al. Heralded entanglement between solid-state qubits separated by three metres. *Nature* **497**, 86–90 (2013).
21. Matsukevich, D. N. et al. Entanglement of remote atomic qubits. *Phys. Rev. Lett.* **96**, 030405 (2006).
22. Ritter, S. et al. An elementary quantum network of single atoms in optical cavities. *Nature* **484**, 195–200 (2012).
23. Axline, C. et al. On-demand quantum state transfer and entanglement between remote microwave cavity memories. *Nat. Phys.* <https://doi.org/10.1038/s41567-018-0115-y> (2018).
24. Campagne-Ibarcq, P. et al. Deterministic Remote Entanglement of Superconducting Circuits through Microwave Two-Photon Transitions. *Phys. Rev. Lett.* **120**, 200501 (2018).
25. Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).
26. Reed, M. D. et al. Fast reset and suppressing spontaneous emission of a superconducting qubit. *Appl. Phys. Lett.* **96**, 203110 (2010).
27. Walter, T. et al. Rapid, high-fidelity, single-shot dispersive readout of superconducting qubits. *Phys. Rev. Appl.* **7**, 054020 (2017).
28. Zeytinoglu, S. et al. Microwave-induced amplitude- and phase-tunable qubit-resonator coupling in circuit quantum electrodynamics. *Phys. Rev. A* **91**, 043846 (2015).
29. Kuhn, A., Hennrich, M. & Rempe, G. Deterministic single-photon source for distributed quantum networking. *Phys. Rev. Lett.* **89**, 067901 (2002).
30. Magnard, P. et al. Fast and unconditional all-microwave reset of a superconducting qubit. Preprint at <https://arxiv.org/abs/1801.07689> (2018).
31. Bozyigit, D. et al. Antibunching of microwave-frequency photons observed in correlation measurements using linear detectors. *Nat. Phys.* **7**, 154–158 (2011).
32. van Enk, S. J., Lütkenhaus, N. & Kimble, H. J. Experimental procedures for entanglement verification. *Phys. Rev. A* **75**, 052318 (2007).
33. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* 10th edn, Ch. 9 (Cambridge Univ. Press, New York, 2011).

**Acknowledgements** This work was supported by the European Research Council (ERC) through the ‘Superconducting Quantum Networks’ (SuperQuNet) project, by the National Centre of Competence in Research ‘Quantum Science and Technology’ (NCCR QSIT), a research instrument of the Swiss National Science Foundation (SNSF), by ETH Zurich and NSERC, the Canada First Research Excellence Fund and the Vanier Canada Graduate Scholarships.

**Author contributions** The experiment was designed and developed by P.K., T.W., P.M. and M.P. The samples were fabricated by J.-C.B., T.W. and S.G. The experiments were performed by P.K., P.M. and T.W. The data were analysed and interpreted by P.K., P.M., B.R., A.B. and A.W. The FPGA firmware and experiment automation was implemented by J.H., Y.S., A.A., S.S., P.M. and P.K. The master-equation simulations were performed by B.R., M.P., P.M. and P.K. The manuscript was written by P.K., P.M., T.W., B.R. and A.W. All authors commented on the manuscript. The project was led by A.W.

**Competing interests** The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0195-y>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to P.K. or A.W. **Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## METHODS

**Literature overview.** In Extended Data Fig. 1, we present a short overview of remote-entanglement experiments performed using the systems and schemes listed in the caption. We calculate bounds<sup>34</sup> on the concurrence  $\mathcal{C}$  for papers in which a CHSH-Bell correlation  $S$  was specified but no value for  $\mathcal{C}$  was given, to provide a more complete comparison. However, we do not calculate  $\mathcal{C}$  for papers that provide only a Bell-state fidelity because knowledge of the elements of the density matrix are necessary to determine  $\mathcal{C}$  without further assumptions.

In addition to the work described here, two independently performed experiments<sup>23,24</sup> concurrently realized deterministic remote state transfer and remote-entanglement generation with superconducting circuits along the lines of a previous proposal<sup>25</sup>. In contrast to our work, the other experiments<sup>23,24</sup> use three-dimensional cavities and transmon qubits with superior coherence properties instead of planar two-dimensional systems. They also use radiation fields with a Gaussian profile and a duration of 2–4  $\mu$ s for state transfer and remote-entanglement generation—substantially longer than used here—which made use of dedicated Purcell filters to increase the emission and absorption bandwidth of the fields used for state transfer. As a result, the concurrence and fidelity found in our experiments exceed those found in the experiments in refs<sup>23,24</sup> at comparable absorption efficiencies and despite the inferior coherence times. Going beyond our work and that presented in ref.<sup>24</sup>, ref.<sup>23</sup> presents transfer of multi-photon states and discusses the potential of the scheme for implementing error correction at each site.

We also note a measurement-based probabilistic realization of remote entanglement in superconducting circuits<sup>35</sup> that improves on previous work<sup>14</sup> and a recent experiment with nitrogen-vacancy centres<sup>36</sup> using single-photon interference and detection to guarantee deterministic delivery of entangled states at a specified time. **Sample parameters.** The designs are very similar to those used previously<sup>27</sup>, with only minor parameter modifications. The  $\lambda/4$  coplanar waveguide resonators and feed lines are created from etched niobium on a sapphire substrate using standard photolithography techniques (Extended Data Fig. 2a) We define the transmon capacitor pads and junctions with electron-beam lithography and shadow-evaporated aluminium with lift-off. We extract the parameters of the readout circuit (grey elements, Fig. 1b) and transfer circuit (yellow elements, Fig. 1b), as well as the coupling strength of the transmon to these circuits, from fits to the transmission spectra of the respective Purcell filter when the transmon is prepared in its ground or excited state using a technique and model discussed previously<sup>27,37,38</sup>. We obtained four working samples from two fabrication runs, with a standard deviation of approximately 8 MHz in the frequency  $\nu_T$  of the transfer resonators, and used the pair with the best-matching frequencies. We then tuned the transfer resonators into resonance ( $\Delta\nu_T \approx 0.2$  MHz) using the dependence of the resonator dispersive shift<sup>39</sup> on the transmon-resonator detuning  $\delta = \nu_{ge} - \nu_T$ . To tune the transmon frequencies we use a miniature superconducting coil to thread flux through the superconducting quantum-interference device (SQUID) at each node. Furthermore, the anharmonicity  $\alpha$  and the coherence times  $T_{2ge}^R$  and  $T_{2ef}^R$  of the qutrits are determined using Ramsey-type measurements. We obtain  $T_{1ef}^R \approx T_{1ge}/3$  for the energy decay times  $T_{1ge}$  and  $T_{1ef}$  of both transmons, which is lower than the expected<sup>40</sup>  $T_{1ge}/2$ . The excess decay rate may be caused by the more complicated environmental mode structure presented to our transmons due to the set of two resonators with their respective Purcell filters coupled to it. All relevant device parameters are listed in Extended Data Table 1.

**Microwave drive schemes.** We use resonant Gaussian DRAG<sup>41,42</sup> microwave pulses of length 19.8 ns and 16.8 ns for  $R_{ge}^\pi$  and  $R_{ef}^\pi$  to swap populations between the  $|g\rangle$  and  $|e\rangle$  states and the  $|e\rangle$  and  $|f\rangle$  states, respectively. We extract an averaged Clifford-gate fidelity for the  $|g\rangle$  and  $|e\rangle$  pulses of more than 99.2% for both transmon qubits, from randomized benchmarking experiments<sup>43</sup>.

We induce the effective coupling  $\tilde{g}$  between states  $|f, 0\rangle$  and  $|g, 1\rangle$  by applying a microwave tone with drive amplitude  $\varepsilon$  to the transmon, at the resonance frequency of the transition ( $\nu_{f0g1}^A = 4.022$  GHz and  $\nu_{f0g1}^B = 3.485$  GHz). Following a procedure described previously<sup>9,30</sup>, we calibrate the a.c. Stark shift of the transmon levels induced by the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  drive, and extract the linear relation between the drive amplitude  $\varepsilon$  and the effective coupling  $\tilde{g}$  (Extended Data Fig. 3). To remain in resonance with the driven transition, we adjust the phase of  $\varepsilon$  on the basis of the measured a.c. Stark shift. We calibrate the drive to reach maximum effective couplings of  $\tilde{g}_m^A/(2\pi) = 6.0$  MHz and  $\tilde{g}_m^B/(2\pi) = 6.7$  MHz (Extended Data Fig. 3b).

We generate photons with temporal shape  $\phi(t) = \frac{1}{2} \sqrt{\kappa_{eff}} \text{sech}(\kappa_{eff}t/2)$  by resonantly driving the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition with

$$\tilde{g}(t) = \frac{\kappa_{eff}}{4 \cosh(\kappa_{eff}t/2)} \frac{1 - e^{\kappa_{eff}t} + (1 + e^{\kappa_{eff}t})\kappa_T/\kappa_{eff}}{\sqrt{(1 + e^{\kappa_{eff}t})\kappa_T/\kappa_{eff} - e^{\kappa_{eff}t}}} \quad (1)$$

where  $\kappa_T$  is the bandwidth of the transfer resonator and  $\kappa_{eff} \leq \kappa_T$  is determined by the strength and duration of the transfer pulse. The dynamics are well described by a two-level model with loss, captured by the non-Hermitian Hamiltonian

$$H = \begin{bmatrix} 0 & \tilde{g} \\ \tilde{g}^* & -i\kappa_T/2 \end{bmatrix} \quad (2)$$

where  $\tilde{g}^*$  is the complex conjugate of  $\tilde{g}$ . This Hamiltonian acts on states  $|f, 0\rangle$  and  $|g, 1\rangle$ , analysed in a rotating frame. The non-Hermitian term  $-i\kappa_T/2$  accounts for photon emission, which brings the system to the dark state  $|g, 0\rangle$ . It can be shown that using the effective coupling of equation (1) in the Hamiltonian in equation (2) leads to the emission of a single photon with the desired temporal shape. This analytical solution provides the option of adjusting the effective bandwidth  $\kappa_{eff}$  of the emitted photon and of generating photon shapes with exponential falling and rising edges at rate  $\kappa_{eff}$ . In all experiments, we create photons with the maximum bandwidth achievable in our set-up, limited by  $\kappa_T$  of node A ( $\kappa_T^A < \kappa_T^B$ ,  $2\tilde{g}_m^A, 2\tilde{g}_m^B$ ).  $\kappa_T^A = \kappa_{eff}$  results in a symmetric amplitude and phase profile of the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transfer pulse at node A and  $\kappa_T^B > \kappa_{eff}$  in an asymmetric drive shape at node B (equation (1)). For the absorption process of the photon we time-reverse the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  drive at node B (Fig. 2g). The photon shape with the shortest pulse duration would require an exponential rising edge proportional to the bandwidth of the receiving node ( $\kappa_T^B$ ) and a falling edge proportional to the bandwidth of the emitting node ( $\kappa_T^A$ ). In our experiment, we create photons with a symmetric shape, approximately realizing the shortest photon duration.

An alternative protocol to generate a remote-entangled state involves preparing the transmon at node A in  $|f, 0\rangle$ , swapping half of the population to  $|g, 1\rangle$  ( $R_{f0g1}^\pi/2$ ) and using the same  $|g, 1\rangle \leftrightarrow |f, 0\rangle$  absorption pulse at node B as actually realized in our experiment (Fig. 4a).  $R_{f0g1}^\pi/2$  can be used to decrease the emission time. However, the absorption process requires the same time. Therefore, in our realization, there is no advantage to using this modified protocol.

**Three-level single-shot readout.** The state of transmon A (B) is read out with a gated microwave tone applied to the input port of the readout resonator Purcell filter at frequency  $\nu_d^A = 4.778$  GHz ( $\nu_d^B = 4.765$  GHz). As depicted in Fig. 1b, the output signal is routed through a set of two circulators and a combiner and then amplified at 10 mK with 22 dB (19.3 dB) gain using a Josephson parametric amplifier (JPA). The JPA pump tone is detuned by 2 MHz from the measurement signal and has a bandwidth of 18.3 MHz (32 MHz). Using these JPAs we find a phase-preserving detection efficiency of  $\eta = 0.61$  ( $\eta = 0.60$ ) for the full detection line. The signal is then further amplified by a high-electron-mobility transistor (HEMT) at 4 K and two low-noise amplifiers at room temperature. Subsequently, the signal is down-converted to 250 MHz using an analogue mixer, lowpass-filtered, digitized by an analogue-to-digital converter and processed by a field-programmable gate array (FPGA). Within the FPGA, the data are digitally down-converted to d.c. and the corresponding I and Q quadrature values are recorded during a window of 256 ns in 8-ns time steps. The FPGA trigger is timed so that the integration window starts with the rising edge of the measurement pulse. We refer to a recording of the I and Q quadrature of a measurement pulse as a readout trace,  $S(t)$ .

We prepare the transmon in states  $|g\rangle$ ,  $|e\rangle$  and  $|f\rangle$  25,000 times each and record the single-shot traces. Each trace is then integrated in post-processing, with two weight functions,  $w_1(t)$  and  $w_2(t)$ , to obtain the integrated quadratures

$$u = \int S(t)w_1(t)dt \quad \text{and} \quad v = \int S(t)w_2(t)dt$$

The collected and integrated traces form three Gaussian-shaped clusters in the  $u$ - $v$  plane (Extended Data Fig. 4), which correspond to the Gaussian probability distributions of the trace when the qutrit is prepared in one of the three eigenstates. We model the probability distribution  $\mathbf{x} = (u, v)$  as a sum of three Gaussian distributions, with density

$$f(\mathbf{x}) = \sum_s \frac{A_s}{2\pi\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_s)^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_s)\right]$$

and estimate the parameters  $A_s$ ,  $\boldsymbol{\mu}_s$  and  $\Sigma$ . On the basis of these parameters, we divide the  $u$ - $v$  plane into the three regions used to assign the result of the readout of the qutrit state (Extended Data Fig. 4). If an integrated trace is in the region labelled  $s'$ , then we assign it state  $s'$ . By counting the number of traces prepared in state  $|s\rangle$  and assigned the value  $s'$ , we estimate the assignment probabilities  $R_{ss'}$  ( $P(s'|s)$ ) (Extended Data Fig. 4). We optimize the measurement power and signal integration time to minimize the measurement error probability  $\|I - R\|_1/6$ , that is, the sum of the off-diagonal elements of the assignment probability matrix (Extended Data Table 2) divided by the number of preparation states. The minimal measurement error probability is realized with an integration time of  $t_m^A = 112$  ns ( $t_m^B = 216$  ns) and a measurement power that results in a state-dependent photon number in the readout resonator between 0.1 and 2. The probabilities of correct assignment range from 93% to 98% for both qutrits (diagonal elements of Extended Data Table 2).

The probability  $M_{s'}$  to assign value  $s'$  to a single-shot measurement of a qutrit in state  $\rho$  is

$$M_{s'} = P(s' | \rho) = \sum_s P(s' | |s\rangle) \rho_{ss}$$

which can be expressed as  $\mathbf{M} = R\rho_{\text{diag}}$  where  $\rho_{\text{diag}}$  is a vector consisting of the diagonal elements of  $\rho$ . The assignment probabilities  $\mathbf{M}$  are typically estimated from assignment counts and a first approach to estimate  $\rho_{\text{diag}}$  is to equate it to  $\mathbf{M}$ . This approach is sensitive to measurement errors, but insensitive to state preparation errors. Setting  $\rho_{\text{diag}} = R^{-1}\mathbf{M}$  effectively accounts for the effect of single-shot readout error. However, this approach relies on the ability to estimate  $R$  precisely and is therefore sensitive to state-preparation error. With transmon reset infidelities of approximately 0.2%<sup>30</sup> and single-qubit gate errors of 0.8% (measured with randomized benchmarking), state-preparation errors are expected to be lower than readout errors. For this reason, we chose to use the latter approach.

We note that the assignment probability matrix  $R_{s_A s_B s'_A s'_B} = P(s'_A s'_B | s_A s_B) = P(s'_A | s_A) P(s'_B | s_B)$  can be obtained as the outer product of the single-qutrit assignment probability matrices (compiled in Extended Data Table 3) and that we can extend this formalism to correct for single-shot readout errors and extract the state populations of a two-qutrit system.

**Estimate of photon loss.** We determine the photon loss  $I_{AB}$  between node A and node B by emitting a photon in the coherent state  $(|0\rangle + |1\rangle)/\sqrt{2}$  first from node A and then from node B independently, as discussed in the main text. Making use of the circulator between the two nodes, we detect the field of each of the emitted photons in the same detection line (Fig. 1). The path travelled by the two emitted photons towards the detector differs only by the length of the waveguide separating the two samples from each other. We evaluate the ratio of the integrated power of the detected fields

$$\frac{\int |\langle a_{\text{out}}^A(t) \rangle|^2 dt}{\int |\langle a_{\text{out}}^B(t) \rangle|^2 dt}$$

to extract the photon loss  $I_{AB}$  between node A and node B. In addition, we estimate the photon loss between node A and B from the specifications of the individual elements connecting the nodes: two printed circuit boards, including connectors (each  $2.5\% \pm 1\%$ ), two coaxial cables of length 0.4 m (each  $4.0\% \pm 0.1\%$ )<sup>44</sup> and a microwave circulator ( $13\% \pm 2\%$  according to manufacturer). With these parameters we estimate an accumulated photon loss between node A and node B of  $24\% \pm 3\%$ , in good agreement with the measured value of  $I_{AB} = 23.0\% \pm 0.5\%$ .

**Master-equation simulation.** We model the transmons as anharmonic oscillators with annihilation (creation) operators  $\hat{b}_i$  ( $\hat{b}_i^\dagger$ )<sup>10</sup>, where the subscript  $i \in \{A, B\}$  denotes the emitter and receiver samples, respectively. The transfer resonator annihilation (creation) operators are denoted as  $\hat{a}_i$  ( $\hat{a}_i^\dagger$ ). Setting  $\hbar = 1$ , the driven Jaynes–Cummings Hamiltonian for sample  $i$  is

$$\begin{aligned} \hat{H}^i = & \omega_T^i \hat{a}_i^\dagger \hat{a}_i + \omega_{\text{ge}}^i \hat{b}_i^\dagger \hat{b}_i + \Omega^i(t) (\hat{b}_i + \hat{b}_i^\dagger) \\ & + g_T^i (\hat{a}_i^\dagger \hat{b}_i + \hat{a}_i \hat{b}_i^\dagger) - \frac{E_C^i}{2} \hat{b}_i^\dagger \hat{b}_i \hat{b}_i^\dagger \hat{b}_i \end{aligned} \quad (3)$$

where  $g_T^i$  denotes the coupling between the transmon and the transfer resonator,  $E_C^i$  denotes the charging energy of the transmon and  $\Omega^i(t) = \Omega^i \cos[\omega_d^i t + \varphi^i(t)]$  is the amplitude of the microwave drive that induces the desired coupling  $\tilde{g}(t)$ . Because the readout resonators do not play a part in the photon transfer dynamics, they are omitted from the Hamiltonian; the static Lamb shifts that they induce are implicitly included in the parameters.

To make the effective coupling  $\tilde{g}(t)$  between the  $|f, 0\rangle$  and  $|g, 1\rangle$  states apparent and to simplify the simulations, we perform a series of unitary transformations on equation (3). We first move to a frame rotating at the drive frequency  $\omega_d^i$ , and then perform a displacement transformation  $\hat{b}_i \rightarrow \hat{b}_i - \beta^i$ ,  $\hat{a}_i \rightarrow \hat{a}_i - \gamma^i$ , choosing  $\beta^i$  and  $\gamma^i$  so that the amplitude of the linear drive terms is set to zero. Next, we perform a Bogoliubov transformation  $\hat{b}_i \rightarrow \cos(A^i) \hat{b}_i - \sin(A^i) \hat{a}_i$ ,  $\hat{a}_i \rightarrow \cos(A^i) \hat{a}_i + \sin(A^i) \hat{b}_i$  where  $\tan(2A^i) = -2g_T^i / (\omega_T^i - \omega_{\text{ge}}^i + 2E_C^i |\beta^i|)$ . Neglecting small off-resonant terms, we obtain the resulting effective Hamiltonian

$$\begin{aligned} \hat{H}_{\text{eff}}^i = & \Delta_T^i \hat{a}_i^\dagger \hat{a}_i + \Delta_{\text{ge}}^i \hat{b}_i^\dagger \hat{b}_i + \frac{\alpha^i}{2} \hat{b}_i^\dagger \hat{b}_i^\dagger \hat{b}_i \hat{b}_i + \frac{K^i}{2} \hat{a}_i^\dagger \hat{a}_i^\dagger \hat{a}_i \hat{a}_i \\ & + 2\chi_T^i \hat{a}_i^\dagger \hat{a}_i \hat{b}_i^\dagger \hat{b}_i + \frac{1}{\sqrt{2}} (\tilde{g} \hat{b}_i^\dagger \hat{b}_i^\dagger \hat{a}_i + \tilde{g}^* \hat{a}_i^\dagger \hat{b}_i \hat{b}_i) \end{aligned} \quad (4)$$

where  $\alpha^i = -E_C^i \cos^4 A^i$  is the transmon anharmonicity,  $K^i = -E_C^i \sin^4 A^i$  is the qubit-induced resonator anharmonicity,  $\chi_T^i = -E_C^i \cos^2 A^i \sin^2 A^i$  is the dispersive shift,  $\Delta_T^i = \omega_T^i \cos^2 A^i + (\omega_{\text{ge}}^i - 2E_C^i |\beta^i|) \sin^2 A^i - g_T^i \sin 2A^i - \omega_d^i$  is the resona-

tor-drive detuning and  $\Delta_{\text{ge}}^i = (\omega_{\text{ge}}^i - 2E_C^i |\beta^i|) \cos^2 A^i + \omega_T^i \sin^2 A^i + g_T^i \sin 2A^i - \omega_d^i$  is the qubit-drive detuning. In equation (4), the desired effective coupling  $\tilde{g}^i = -E_C^i \beta^i \sqrt{2} \cos^2 A^i \sin A^i$  between the  $|f, 0\rangle$  and  $|g, 1\rangle$  states is now made explicit.

Finally, moving to a frame rotating at  $\Delta_T^i$  for the resonator and  $\Delta_{\text{ge}}^i + \alpha^i/2$  for the transmon qubits, the combined effective Hamiltonian of the two samples is

$$\begin{aligned} \hat{H}_{\text{eff}} = & \sum_{i=A,B} \left\{ -\frac{\alpha^i}{2} \hat{b}_i^\dagger \hat{b}_i + \frac{\alpha^i}{2} \hat{b}_i^\dagger \hat{b}_i^\dagger \hat{b}_i \hat{b}_i \right. \\ & + \frac{K^i}{2} \hat{a}_i^\dagger \hat{a}_i^\dagger \hat{a}_i \hat{a}_i + 2\chi_T^i \hat{a}_i^\dagger \hat{a}_i \hat{b}_i^\dagger \hat{b}_i \\ & \left. + \frac{1}{\sqrt{2}} [\tilde{g}^i(t) \hat{b}_i^\dagger \hat{b}_i^\dagger \hat{a}_i + \tilde{g}^i(t)^* \hat{a}_i^\dagger \hat{b}_i \hat{b}_i] \right\} \\ & - i \frac{\sqrt{\kappa_T^A \kappa_T^B} \eta_c}{2} (\hat{a}_A \hat{a}_B^\dagger - \hat{a}_A^\dagger \hat{a}_B) \end{aligned}$$

where  $\eta_c$  is the photon-loss probability of the circulator between the two samples.

Using this effective Hamiltonian, numerical results are obtained by integrating the master equation

$$\begin{aligned} \dot{\rho} = & -i[\hat{H}_{\text{eff}}, \rho] \\ & + \kappa_T^A (1 - \eta_c) \mathcal{D}[\hat{a}_A] \rho + \mathcal{D}[\sqrt{\kappa_T^A} \eta_c \hat{a}_A + \sqrt{\kappa_T^B} \hat{a}_B] \rho \\ & + \sum_{i=A,B} \{ \kappa_{\text{int}}^i \mathcal{D}[\hat{a}_i] \rho + \gamma_{\text{ig}}^i \mathcal{D}[|g\rangle \langle e|_i] \rho \\ & + \gamma_{\text{if}}^i \mathcal{D}[|e\rangle \langle f|_i] \rho \} \\ & + \sum_{i=A,B} \{ \gamma_{\phi_{\text{ge}}}^i \mathcal{D}[|e\rangle \langle e|_i - |g\rangle \langle g|_i] \rho \\ & + \gamma_{\phi_{\text{ef}}}^i \mathcal{D}[|f\rangle \langle f|_i - |e\rangle \langle e|_i] \rho \} \end{aligned} \quad (5)$$

where  $\mathcal{D}[\hat{O}] \bullet = \hat{O} \bullet \hat{O}^\dagger - \{\hat{O}^\dagger \hat{O}, \bullet\}/2$  denotes the dissipation super-operator,  $\kappa_{\text{int}}^i$  the internal decay rates of the resonators,  $\gamma_{\text{inm}}^i = 1/T_{1nm}^i$  the decay rates of the transmon qubits between the  $|n\rangle_i$  and  $|m\rangle_i$  states and  $\gamma_{\phi_{nm}}^i = 1/(2T_{1nm}^i) - 1/(T_{2nm}^i)$  the dephasing rates between the  $|n\rangle_i$  and  $|m\rangle_i$  states of the transmon qubits. The last term in  $\hat{H}_{\text{eff}}$  combined with the resonator dissipators in the second line of the master equation (equation (5)), assure that the output of the emitter A is cascaded to the input of the receiver B<sup>45,46</sup> through a circulator with photon loss  $\eta_c$ .

**Quantum state and process tomography.** Quantum state tomography of a single qutrit is performed by measuring the qutrit state population with the single-shot readout method, after applying the following tomography gates:  ${}^x R_{\text{ge}}^0$ ,  ${}^x R_{\text{ge}}^{\pi/2}$ ,  ${}^y R_{\text{ge}}^{\pi/2}$ ,  ${}^x R_{\text{ef}}^{\pi/2}$ ,  ${}^y R_{\text{ef}}^{\pi/2}$ ,  $({}^x R_{\text{ge}}^{\pi} {}^x R_{\text{ef}}^{\pi/2})$ ,  $({}^x R_{\text{ge}}^{\pi} {}^y R_{\text{ef}}^{\pi/2})$  and  $({}^x R_{\text{ge}}^{\pi} {}^x R_{\text{ef}}^{\pi})$ . The elements of the density matrix are then reconstructed using a maximum-likelihood method, assuming ideal tomography gates.

To extend this quantum state tomography procedure to two-qutrit density matrices, we perform two local tomography gates (from the 81 pairs of gates that can be formed from the single-qutrit quantum state tomography gates) on transmons A and B, before extracting the state populations using the two-qutrit single-shot measurement method.

To characterize the qubit state transfer from node A to node B, we perform full quantum process tomography<sup>47</sup> (Fig. 3, Extended Data Table 4). We prepare each of the six mutually unbiased qubit basis states<sup>32</sup>  $|g\rangle$ ,  $|e\rangle$ ,  $(|g\rangle + |e\rangle)/\sqrt{2}$ ,  $(|g\rangle + i|e\rangle)/\sqrt{2}$ ,  $(|g\rangle - |e\rangle)/\sqrt{2}$  and  $(|g\rangle - i|e\rangle)/\sqrt{2}$ , transfer the state to node B, then independently measure the qutrit density matrix at node A and node B with quantum state tomography. We obtain the process matrix from these density matrices through linear inversion. Quantum state tomography of the qutrit subspace is required to characterize the residual population in  $|f\rangle$  after the qubit state transfer, which is caused mainly by decay from the  $|f\rangle$  level in combination with  $R_{\text{ef}}^{\pi}$  swapping  $|e\rangle$  with  $|f\rangle$  populations. The density matrices that we obtain have a non-unit trace in the qubit subspace, and so does the qubit state transfer process matrix. Consequences of that observation are discussed below for the example of the Bell-state density matrix.

**Two-qutrit entanglement.** Owing to a residual population of 3.5% of the  $|f\rangle$  level of the transmons after the entanglement protocol, the entangled state cannot be described rigorously by a two-qubit density matrix. To be concise, we represent the reconstructed two-qutrit entangled state  $\rho_{3\otimes 3}$  (Extended Data Fig. 5, Extended Data Table 5) by a two-qubit density matrix  $\rho_m$  that consists of the two-qubit elements of  $\rho_{3\otimes 3}$ . This choice of reduction from a two-qutrit to a two-qubit density matrix conserves the state fidelity  $\mathcal{F}_{|\psi^+\rangle} = \langle \psi^+ | \rho_m | \psi^+ \rangle = \langle \psi^+ | \rho_{3\otimes 3} | \psi^+ \rangle$ ; however,  $\rho_m$  has a non-unit trace. In addition, this reduction method gives a

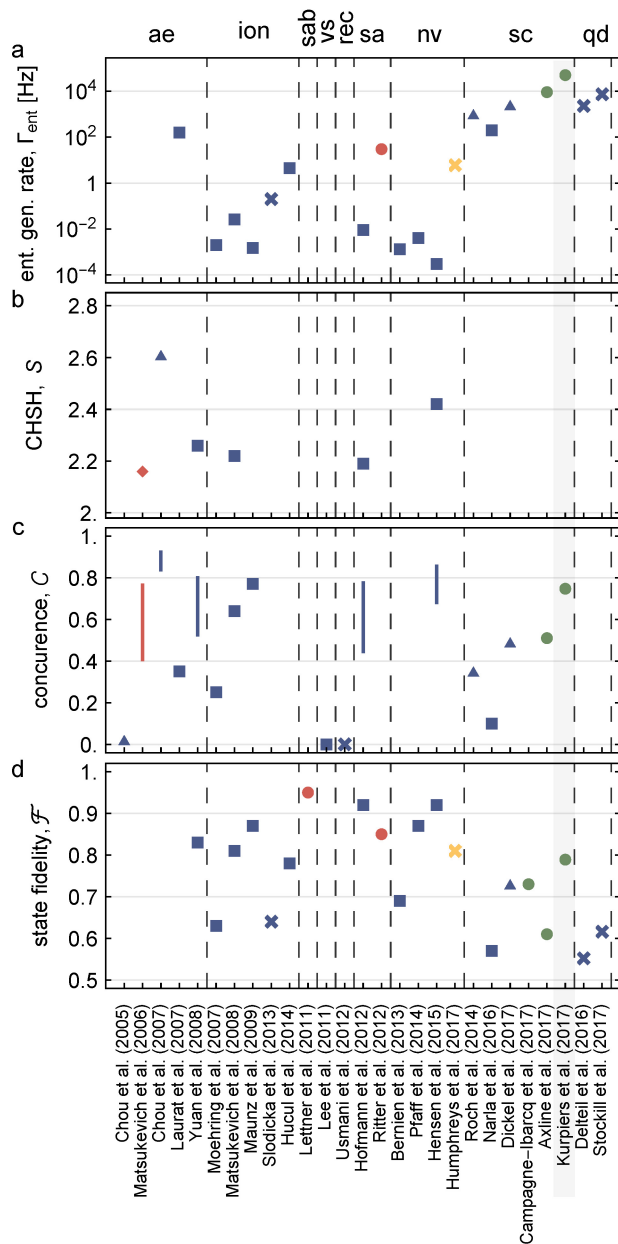
conservative estimate of the concurrence  $\mathcal{C}(\rho_m)$  compared to a projection of  $\rho_{3\otimes 3}$  on the set of physical two-qubit density matrices.

To verify the three-level bipartite entanglement, we use the computable cross-norm or realignment criterion<sup>48</sup>, which is well defined for multi-level mixed entangled states. This criterion states that a state  $\rho$  must be entangled if  $\sum_k \lambda_k > 1$ , with  $\rho = \sum_k \lambda_k G_k^A \otimes G_k^B$  and  $G_k^{A(B)}$  an orthonormal basis of the observable spaces of  $\mathcal{H}^{A(B)}$ . We obtain  $\sum_k \lambda_k = 1.612 \pm 0.003$  with the measured entangled state  $\rho_{3\otimes 3}$ , providing unambiguous evidence for the existence of entanglement of the prepared state.

**Data availability.** The data that support the findings of this study are available within the paper.

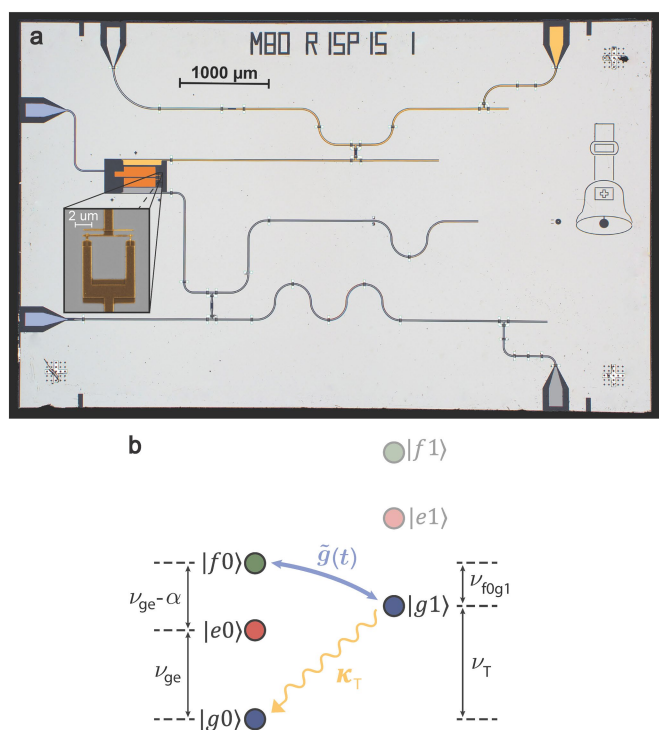
34. Verstraete, F. & Wolf, M. M. Entanglement versus bell violations and their behavior under local filtering operations. *Phys. Rev. Lett.* **89**, 170401 (2002).
35. Dickel, C. et al. Chip-to-chip entanglement of transmon qubits using engineered measurement fields. *Phys. Rev. B* **97**, 064508 (2018).
36. Humphreys, P. C. et al. Deterministic delivery of remote entanglement on a quantum network. Preprint at <https://arxiv.org/abs/1712.07567> (2017).
37. Jeffrey, E. et al. Fast accurate state measurement with superconducting qubits. *Phys. Rev. Lett.* **112**, 190504 (2014).
38. Sete, E. A., Martinis, J. M. & Korotkov, A. N. Quantum theory of a bandpass Purcell filter for qubit readout. *Phys. Rev. A* **92**, 012325 (2015).
39. Wallraff, A. et al. Approaching unit visibility for control of a superconducting qubit with dispersive readout. *Phys. Rev. Lett.* **95**, 060501–060504 (2005).
40. Peterer, M. J. et al. Coherence and decay of higher energy levels of a superconducting transmon qubit. *Phys. Rev. Lett.* **114**, 010501 (2015).
41. Motzoi, F., Gambetta, J. M., Rebentrost, P. & Wilhelm, F. K. Simple pulses for elimination of leakage in weakly nonlinear qubits. *Phys. Rev. Lett.* **103**, 110501 (2009).
42. Gambetta, J. M., Motzoi, F., Merkel, S. T. & Wilhelm, F. K. Analytic control methods for high-fidelity unitary operations in a weakly nonlinear oscillator. *Phys. Rev. A* **83**, 012308–012313 (2011).
43. Chow, J. M. et al. Randomized benchmarking and process tomography for gate errors in a solid-state qubit. *Phys. Rev. Lett.* **102**, 090502 (2009).
44. Kurpiers, P., Walter, T., Magnard, P., Salathe, Y. & Wallraff, A. Characterizing the attenuation of coaxial and rectangular microwave-frequency waveguides at cryogenic temperatures. *EPJ Quant. Technol.* **4**, 8 (2017).
45. Gardiner, C. W. Driving a quantum system with the output field from another driven quantum system. *Phys. Rev. Lett.* **70**, 2269–2272 (1993).
46. Carmichael, H. J. Quantum trajectory theory for cascaded open systems. *Phys. Rev. Lett.* **70**, 2273–2276 (1993).
47. Chuang, I. L. & Nielsen, M. A. Prescription for experimental determination of the dynamics of a quantum black box. *J. Mod. Opt.* **44**, 2455–2467 (1997).
48. Gühne, O. & Tóth, G. Entanglement detection. *Phys. Rep.* **474**, 1–75 (2009).
49. Bell, J. S. On the Einstein Podolsky Rosen paradox. *Phys. Phys. Fiz.* **1**, 195–200 (1964).
50. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
51. Chou, C.-W. et al. Functional quantum nodes for entanglement distribution over scalable quantum networks. *Science* **316**, 1316–1320 (2007).
52. Laurat, J., Choi, K. S., Deng, H., Chou, C. W. & Kimble, H. J. Heralded entanglement between atomic ensembles: preparation, decoherence, and scaling. *Phys. Rev. Lett.* **99**, 180504 (2007).
53. Yuan, Z.-S. et al. Experimental demonstration of a BDCZ quantum repeater node. *Nature* **454**, 1098–1101 (2008).
54. Matsukevich, D. N., Maunz, P., Moehring, D. L., Olmschenk, S. & Monroe, C. Bell inequality violation with two remote atomic qubits. *Phys. Rev. Lett.* **100**, 150404 (2008).
55. Maunz, P. et al. Heralded quantum gate between remote quantum memories. *Phys. Rev. Lett.* **102**, 250502 (2009).
56. Hucul, D. et al. Modular entanglement of atomic qubits using photons and phonons. *Nat. Phys.* **11**, 37–42 (2015).
57. Lettner, M. et al. Remote entanglement between a single atom and a Bose-Einstein condensate. *Phys. Rev. Lett.* **106**, 210503 (2011).
58. Usmani, I. et al. Heralded quantum entanglement between two crystals. *Nat. Photon.* **6**, 234–237 (2012).
59. Pfaff, W. et al. Unconditional quantum teleportation between distant solid-state quantum bits. *Science* **345**, 532–535 (2014).
60. Hensen, B. et al. Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
61. Narla, A. et al. Robust concurrent remote entanglement between two superconducting qubits. *Phys. Rev. X* **6**, 031036 (2016).
62. Stockill, R. et al. Phase-tuned entangled state generation between distant spin qubits. *Phys. Rev. Lett.* **119**, 010503 (2017).





**Extended Data Fig. 1 | Overview of remote-entanglement experiments.**

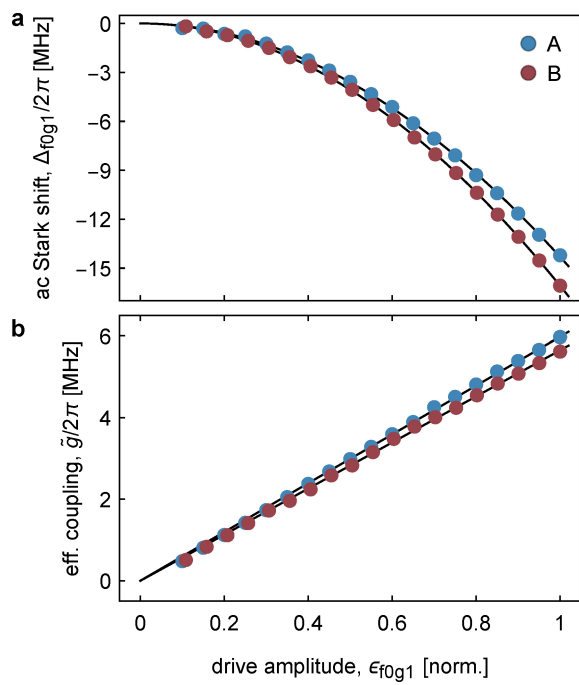
**a**, Entanglement generation rate  $\Gamma_{\text{ent}}$ . **b**, CHSH-Bell inequality<sup>49,50</sup> correlation  $S$ . **c**, Concurrence  $C$ . **d**, Entangled state fidelity  $\mathcal{F}$ . The experiments are grouped by physical system: atomic ensembles ('ae')<sup>12,21,51–53</sup>, trapped ions ('ion')<sup>15,17,54–56</sup>, single-atom Bose-Einstein condensate ('sab')<sup>57</sup>, vibrational state of diamonds ('vs')<sup>18</sup>, rare-Earth-doped crystals ('rec')<sup>58</sup>, single atoms ('sa')<sup>19,22</sup>, nitrogen-vacancy centres ('nv')<sup>20,36,59,60</sup>, superconducting circuits ('sc')<sup>14,23,24,35,61</sup> or quantum dots ('qd')<sup>16,62</sup>. The colours indicate different implementations: probabilistic unheralded (red), probabilistic heralded (blue), guaranteeing a deterministic delivery of an entangled state at a pre-specified time (yellow) or fully deterministic (green). The plot markers indicate different schemes for realizing the remote interaction: measurement-induced (triangles), single-photon (crosses) or two-photon (squares) interference and detection, direct transfer (diamond) or direct transfer with shaped photons (circles). The lines in **c** are bounds<sup>34</sup> on the concurrence  $C$  calculated from measured CHSH-Bell correlations  $S$ . The shaded column highlights this study.



**Extended Data Fig. 2 | Micrograph of sample and energy-level diagram.**

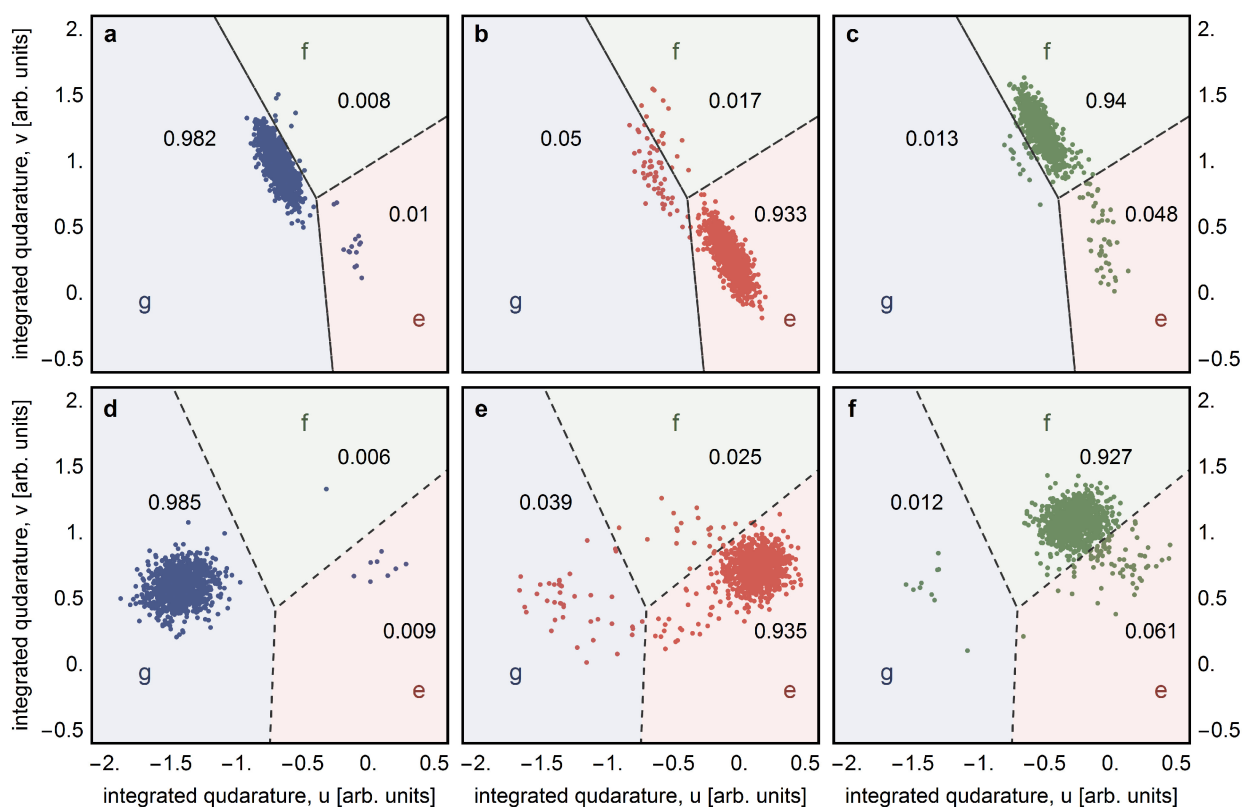
**a**, False-colour micrograph of a sample of the same design co-fabricated with the one used for node A. The circuit elements are colour coded as in Fig. 1: transfer circuit (yellow), readout circuit (grey), transmon (orange) and input lines of the transmon and readout circuit (blue). The input to the transfer circuit is used as an auxiliary port to perform resonator spectroscopy in transmission. The inset shows a scanning electron microscopy (SEM) micrograph of the asymmetric SQUID with a ratio of 5:1 between the areas of the Josephson junctions used in the transmon.

**b**, Schematic of the energy-level diagram of the coupled transmon-transfer resonator system. The numerical values of all parameters are listed in Extended Data Table 1.



**Extended Data Fig. 3 | a.c. Stark shift and Rabi rate of the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition. a, b,** Measurement (filled circles) of the a.c. Stark shift  $\Delta f_{0g1}/(2\pi)$  (a) and the effective coupling  $\tilde{g}/(2\pi)$  of the  $|f, 0\rangle \leftrightarrow |g, 1\rangle$  transition (b) versus drive amplitude  $\epsilon_{f0g1}$  for sample A (blue) and B (red). The solid lines in a (b) are quadratic (linear) fits to the data<sup>30</sup>.

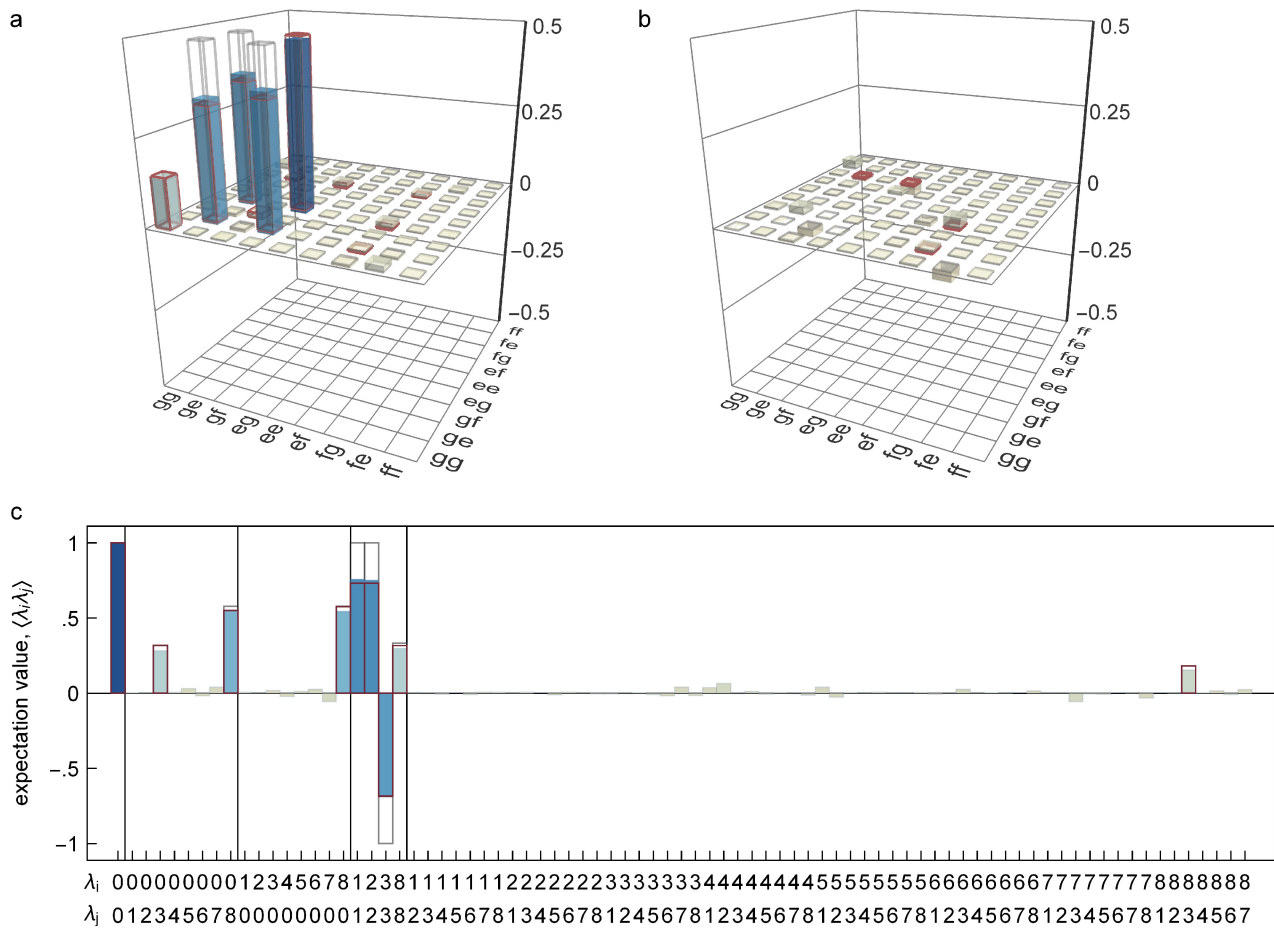




**Extended Data Fig. 4 | Qutrit single-shot readout characterization.**

**a–f**, Scatter plot of the measured integrated quadrature values  $u$  and  $v$  for qutrit A (**a–c**) and B (**d–f**) when prepared in state  $|g\rangle$  (blue; **a**, **d**),  $|e\rangle$  (red; **b**, **e**) and  $|f\rangle$  (green; **c**, **f**). We plot only the first 1,000 of the 25,000

repetitions for each state-preparation experiment. The dashed lines are the qutrit-state discrimination thresholds used to obtain the assignment probabilities (numbers, which are also listed in Extended Data Table 2).



**Extended Data Fig. 5 | Characterization of entangled states in a two-qutrit basis. a–c,** Two-qutrit state tomography: real (a) and imaginary (b) part of the density matrix and expectation values of the Gell–Mann operators  $\lambda_k$  (c). The ideal Bell state  $|\psi^+\rangle$  and numerical master-equation simulation are depicted as grey and red outlines, respectively.  $\lambda_0$  denotes

the identity operation,  $\lambda_{1,2,3}$  denote the Pauli matrices  $\sigma_{x,y,z}^{\text{ge}}$  in the qubit subspace,  $\lambda_{4,5}$  correspond to  $\sigma_{x,y}^{\text{gf}}$ ,  $\lambda_{6,7}$  correspond to  $\sigma_{x,y}^{\text{ef}}$  and  $\lambda_8$  is the diagonal matrix  $(\sigma_z^{\text{ge}} + 2\sigma_z^{\text{ef}}) / \sqrt{3}$ . The trace distance between the measurement and the simulation is 0.107.

**Extended Data Table 1 | Summary of device parameters for nodes A and B**

quantity, symbol (unit)	Node A	Node B
readout resonator frequency, $\nu_R$ (GHz)	4.787	4.780
readout Purcell filter frequency, $\nu_{Rpf}$ (GHz)	4.778	4.780
readout resonator bandwidth, $\kappa_R/2\pi$ (MHz)	12.6	27.1
readout circuit dispersive shift, $\chi_R/2\pi$ (MHz)	5.8	11.6
transfer resonator frequency, $\nu_T$ (GHz)	8.4005	8.4003
transfer Purcell filter frequency, $\nu_{Tpf}$ (GHz)	8.426	8.415
transfer resonator bandwidth, $\kappa_T/2\pi$ (MHz)	10.4	13.5
transfer circuit dispersive shift, $\chi_T/2\pi$ (MHz)	6.3	4.7
qubit transition frequency, $\nu_{ge}$ (GHz)	6.343	6.096
transmon anharmonicity, $\alpha$ (MHz)	-265	-308
$ f, 0\rangle \leftrightarrow  g, 1\rangle$ transition frequency, $\nu_{f0g1}$ (GHz)	4.022	3.485
$ f, 0\rangle \leftrightarrow  g, 1\rangle$ max. eff. coupling, $\tilde{g}_m/2\pi$ (MHz)	6.0	6.7
energy relaxation time on $ge$ , $T_{1ge}$ ( $\mu s$ )	4.9	4.6
energy relaxation time on $ef$ , $T_{1ef}$ ( $\mu s$ )	1.6	1.4
coherence time on $ge$ , $T_{2ge}^R$ ( $\mu s$ )	3.4	2.6
coherence time on $ef$ , $T_{2ef}^R$ ( $\mu s$ )	2.1	0.9



**Extended Data Table 2 | Probabilities of identifying the prepared states (columns) as the measured states (rows) for qutrits A and B**

Qutrit A				Qutrit B			
	$ g\rangle$	$ e\rangle$	$ f\rangle$	$ g\rangle$	$ e\rangle$	$ f\rangle$	
g	98.2	5.0	1.3	98.5	3.9	1.2	g
e	1.0	93.3	4.8	0.9	93.5	6.1	e
f	0.8	1.7	94.0	0.6	2.5	92.7	f

The diagonal elements show correct identification; the off-diagonal elements show misidentifications.

**Extended Data Table 3 | Probabilities of identifying the prepared input states (columns) as the indicated output states (rows) for all possible tuples of two-qutrit basis states**

	$ gg\rangle$	$ ge\rangle$	$ gf\rangle$	$ eg\rangle$	$ ee\rangle$	$ ef\rangle$	$ fg\rangle$	$ fe\rangle$	$ ff\rangle$
gg	96.8	3.9	1.1	4.9	0.2	0.1	1.2	0.0	0.0
ge	0.9	91.9	6.0	0.0	4.7	0.3	0.0	1.2	0.1
gf	0.6	2.5	91.1	0.0	0.1	4.6	0.0	0.0	1.2
eg	1.0	0.0	0.0	91.9	3.7	1.1	4.7	0.2	0.1
ee	0.0	0.9	0.1	0.8	87.3	5.7	0.0	4.5	0.3
ef	0.0	0.0	0.9	0.6	2.4	86.5	0.0	0.1	4.4
fg	0.8	0.0	0.0	1.6	0.1	0.0	92.5	3.7	1.1
fe	0.0	0.7	0.0	0.0	1.6	0.1	0.8	87.9	5.8
ff	0.0	0.0	0.7	0.0	0.0	1.6	0.6	2.4	87.1

The diagonal elements show correct identification; the off-diagonal elements show misidentifications.

**Extended Data Table 4 | Numerical values of the experimentally obtained process-matrix elements of the qubit state transfer**

	I	X	$\tilde{Y}$	Z
I	0.8002	$0.0008-0.0021i$	$-0.0036-0.0033i$	$0.0818-0.0038i$
X	$0.0008+0.0021i$	0.0748	-0.0734	$0.0017+0.0037i$
$\tilde{Y}$	$-0.0036+0.0033i$	-0.0734	0.0727	$0.0011+0.0028i$
Z	$0.0818+0.0038i$	$0.0017-0.0037i$	$0.0011-0.0028i$	0.0313

The absolute value of this process matrix is depicted in Fig. 3 as coloured bars.



**Extended Data Table 5 | Numerical values of the experimentally obtained density-matrix elements of the two-qubit remote-entangled state in a two-qutrit basis**

	gg	ge	gf	eg	ee	ef	fg	fe	ff
gg	0.142	-0.001	0.002-0.005 <i>i</i>	0.001-0.001 <i>i</i>	0.001-0.001 <i>i</i>	0.001	-0.006-0.004 <i>i</i>	0.016-0.027 <i>i</i>	0.001+0.001 <i>i</i>
ge	-0.001	0.343	-0.008-0.021 <i>i</i>	0.378	0.003+0.002 <i>i</i>	-0.003-0.004 <i>i</i>	0.003+0.006 <i>i</i>	-0.005-0.002 <i>i</i>	0.001
gf	0.002+0.005 <i>i</i>	-0.008+0.021 <i>i</i>	0.015	-0.004	-0.002	-0.001-0.002 <i>i</i>	0.005-0.002 <i>i</i>	0.002-0.001 <i>i</i>	0.
eg	0.001+0.001 <i>i</i>	0.378	-0.004	0.48	0.002-0.001 <i>i</i>	0.001-0.01 <i>i</i>	0.013+0.029 <i>i</i>	-0.002-0.001 <i>i</i>	0.002
ee	0.001+0.001 <i>i</i>	0.003-0.002 <i>i</i>	-0.002	0.002+0.001 <i>i</i>	0.	0.	-0.001	0.	0.
ef	0.001	-0.003+0.004 <i>i</i>	-0.001+0.002 <i>i</i>	0.001+0.01 <i>i</i>	0.	0.001	-0.001+0.002 <i>i</i>	0.	0.
fg	-0.006+0.004 <i>i</i>	0.003-0.006 <i>i</i>	0.005+0.002 <i>i</i>	0.013-0.029 <i>i</i>	-0.001	-0.001-0.002 <i>i</i>	0.012	0.001+0.001 <i>i</i>	0.
fe	0.016+0.027 <i>i</i>	-0.005+0.002 <i>i</i>	0.002+0.001 <i>i</i>	-0.002+0.001 <i>i</i>	0.	0.	0.001-0.001 <i>i</i>	0.007	0.
ff	0.001-0.001 <i>i</i>	0.001	0.	0.002	0.	0.	0.	0.	0.

The real and imaginary parts of this density matrix are depicted as coloured bars in Extended Data Fig. 5a and b, respectively.

# Deterministic delivery of remote entanglement on a quantum network

Peter C. Humphreys<sup>1,3</sup>, Norbert Kalb<sup>1,3</sup>, Jaco P. J. Morits<sup>1</sup>, Raymond N. Schouten<sup>1</sup>, Raymond F. L. Vermeulen<sup>1</sup>, Daniel J. Twitchen<sup>2</sup>, Matthew Markham<sup>2</sup> & Ronald Hanson<sup>1\*</sup>

**Large-scale quantum networks promise to enable secure communication, distributed quantum computing, enhanced sensing and fundamental tests of quantum mechanics through the distribution of entanglement across nodes<sup>1–7</sup>. Moving beyond current two-node networks<sup>8–13</sup> requires the rate of entanglement generation between nodes to exceed the decoherence (loss) rate of the entanglement. If this criterion is met, intrinsically probabilistic entangling protocols can be used to provide deterministic remote entanglement at pre-specified times. Here we demonstrate this using diamond spin qubit nodes separated by two metres. We realize a fully heralded single-photon entanglement protocol that achieves entangling rates of up to 39 hertz, three orders of magnitude higher than previously demonstrated two-photon protocols on this platform<sup>14</sup>. At the same time, we suppress the decoherence rate of remote-entangled states to five hertz through dynamical decoupling. By combining these results with efficient charge-state control and mitigation of spectral diffusion, we deterministically deliver a fresh remote state with an average entanglement fidelity of more than 0.5 at every clock cycle of about 100 milliseconds without any pre- or post-selection. These results demonstrate a key building block for extended quantum networks and open the door to entanglement distribution across multiple remote nodes.**

The power of future quantum networks will derive from entanglement that is shared between the network nodes. Two critical parameters for the performance of such networks are the entanglement-generation rate  $r_{\text{ent}}$  between nodes and the entangled-state decoherence rate  $r_{\text{dec}}$ . Their ratio  $\eta_{\text{link}} = r_{\text{ent}}/r_{\text{dec}}$ , which we term the quantum link efficiency<sup>8,15</sup>, quantifies how effectively entangled states can be preserved over the timescales necessary to generate them. Alternatively, the link efficiency determines the average number of entangled states that can be created within one entangled-state lifetime. A link efficiency of unity therefore represents a critical threshold above which entanglement can be generated faster than it is lost. Exceeding this threshold is central to allowing multiple entangled links to be created and maintained simultaneously, as is required for the distribution of many-body quantum states across a network<sup>6,15</sup>.

Consider an elementary entanglement-delivery protocol that delivers states at pre-determined times. This can be achieved by making multiple attempts to generate entanglement and then protecting successfully generated entangled states from decoherence until the required delivery time (Fig. 1a, steps (1)–(3)). If we try to generate entanglement for a period  $t_{\text{ent}}$ , then the cumulative probability of success will be  $p_{\text{succ}} = 1 - e^{-r_{\text{ent}}t_{\text{ent}}}$ . For a given  $p_{\text{succ}}$ , the average fidelity  $F_{\text{succ}}$  with respect to a maximally entangled state of the successfully generated states is solely determined by the quantum link efficiency  $\eta_{\text{link}}$  (Methods). We plot  $F_{\text{succ}}$  versus  $p_{\text{succ}}$  for several values of  $\eta_{\text{link}}$  in Fig. 1b.

This protocol allows entangled states to be delivered at specified times, but with a finite probability of success. By delivering an unentangled state (state fidelity  $F_{\text{unent}} \leq 1/2$ ) in cycles in which all

entanglement-generation attempts failed, the protocol can be cast into a fully deterministic black box (Fig. 1a, step (4)). The states output from such a black box will have a fidelity of

$$F_{\text{det}} = p_{\text{succ}} F_{\text{succ}} + (1 - p_{\text{succ}}) F_{\text{unent}} \quad (1)$$

The maximum achievable fidelity  $F_{\text{det}}^{\text{max}}$  of this deterministic state-delivery protocol, found by optimizing  $p_{\text{succ}}$ , is also determined only by the quantum link efficiency  $\eta_{\text{link}}$ . For  $F_{\text{unent}} = 1/4$  (a fully mixed state), we find (Fig. 1c)

$$F_{\text{det}}^{\text{max}} = \frac{1}{4} [1 + 3\eta_{\text{link}}^{1/(1-\eta_{\text{link}})}] \quad (2)$$

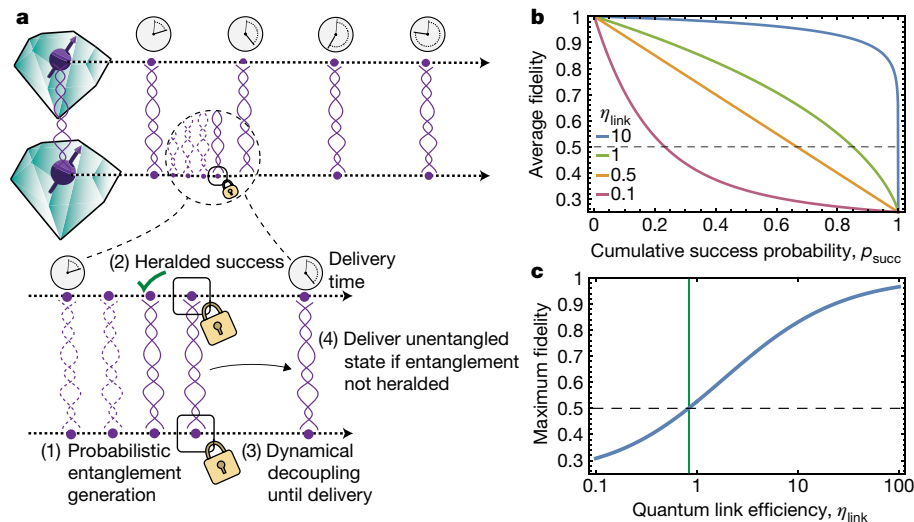
For  $\eta_{\text{link}}$  greater than about 0.83, there exists a combination of  $p_{\text{succ}}$  and  $F_{\text{succ}}$  high enough to compensate for cycles in which entanglement is not heralded, enabling the deterministic delivery of states that are entangled on average ( $F_{\text{det}}^{\text{max}} \geq 1/2$ ). Deterministic entanglement delivery is therefore a critical benchmark of the performance of a network, certifying that its quantum link efficiency is of order unity or higher. Furthermore, the ability to specify in advance the time at which entangled states are delivered may assist in designing multi-step quantum-information tasks such as entanglement routing<sup>16,17</sup>.

However, so far, quantum link efficiencies of order unity or greater have remained out of reach for solid-state quantum networks. Quantum dots have been used to demonstrate kilohertz entanglement rates  $r_{\text{ent}}$ , but decoherence rates  $r_{\text{dec}}$  of tens of megahertz limit their quantum link efficiencies<sup>18,19</sup>  $\eta_{\text{link}}$  to around  $10^{-4}$ . Nitrogen–vacancy (NV) centres—point defects in diamond with a long-lived electron spin and bright optical transitions—have been used to demonstrate entanglement rates  $r_{\text{ent}}$  of tens of millihertz<sup>10,14</sup> and, in separate experiments, decoherence rates  $r_{\text{dec}}$  of the order of one hertz<sup>20</sup>, which together would give link efficiencies  $\eta_{\text{link}}$  of roughly  $10^{-2}$ .

Here we achieve quantum link efficiencies  $\eta_{\text{link}}$  well in excess of unity by realizing an alternative entanglement protocol for NV centres in which we directly use the state heralded by the detection of a single photon (Fig. 2)<sup>21,22</sup>. The rate for such a single-photon protocol scales linearly with losses, which, in comparison with previously used two-photon-mediated protocols<sup>9,14</sup>, provides a substantial advantage in typical remote-entanglement settings. Recent experiments have highlighted the potential of single-photon protocols by generating local entanglement<sup>23,24</sup>, and remote entanglement in post-selection<sup>18,19</sup>. By realizing a single-photon protocol in a fully heralded fashion and protecting entanglement through dynamical decoupling, we achieve the deterministic delivery of remote-entangled states on an approximately 10-Hz clock.

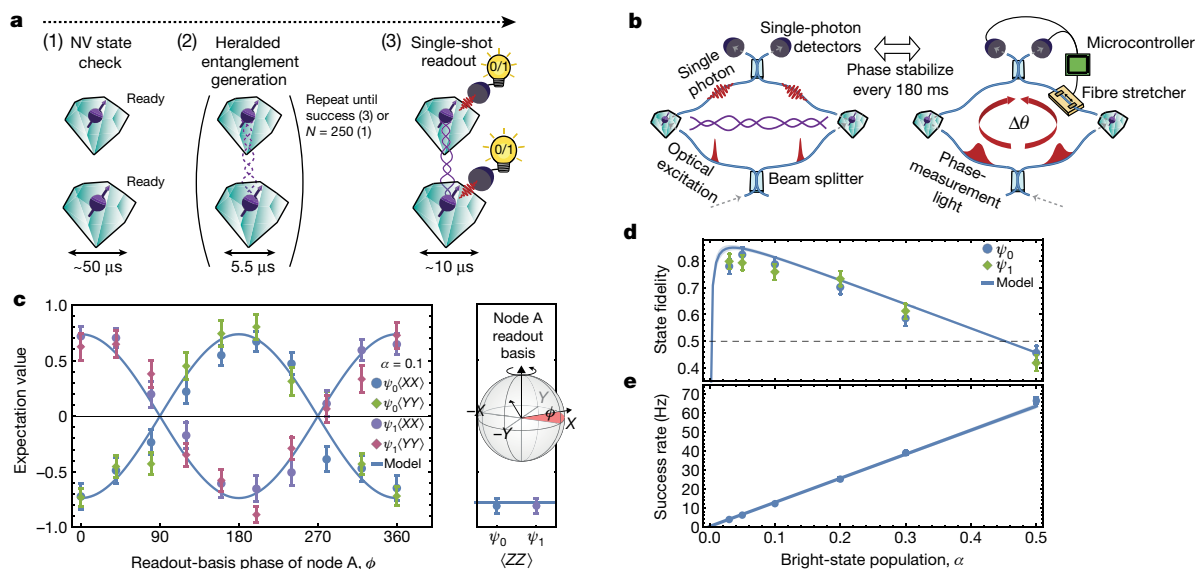
Our experiment uses NV centres that reside in independently operated cryostat set-ups separated by 2 m (further experimental details are given in Methods). We use qubits formed by two of the ground-state spin sublevels of the NV centre ( $|\uparrow\rangle \equiv |m_s = 0\rangle$  and  $|\downarrow\rangle \equiv |m_s = -1\rangle$ , where  $m_s$  is the projection of the spin along its quantisation axis).

<sup>1</sup>QuTech and Kavli Institute of Nanoscience, Delft University of Technology, Delft, The Netherlands. <sup>2</sup>Element Six Innovation, Didcot, UK. <sup>3</sup>These authors contributed equally: Peter C. Humphreys, Norbert Kalb. \*e-mail: r.hanson@tudelft.nl



**Fig. 1 | Deterministic remote-entanglement delivery.** **a**, Deterministic entanglement delivery guarantees the output of states with an average entanglement fidelity of more than 0.5 at pre-specified times. In our protocol, underlying this deterministic delivery is a probabilistic but heralded entanglement process. Repeated entangling attempts (dashed helical links) are made (1) and then, upon heralded success (2), the entangled state (solid helical link) is protected from decoherence (represented by the lock) until the specified delivery time (3). If no attempt at entanglement generation succeeds within one cycle, an unentangled

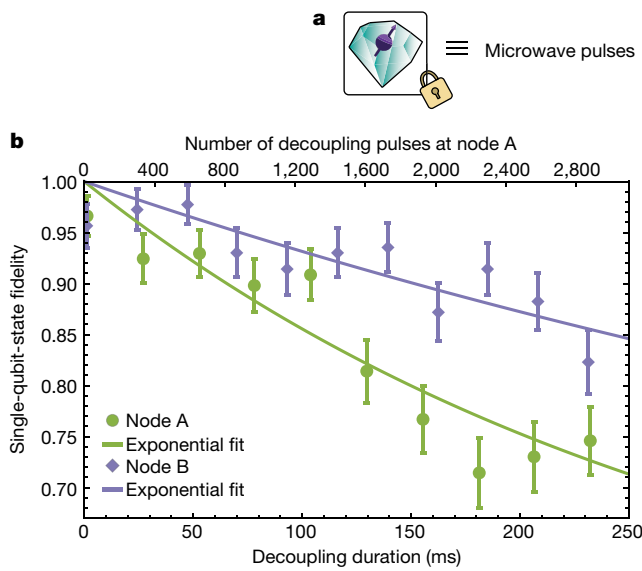
state must be delivered (4). **b**, For the underlying entanglement-generation and state-preservation protocol (steps (1)–(3) in **a**), the effectiveness of the trade-off between the average fidelity of the entangled state that is delivered and the success probability is determined by the quantum link efficiency  $\eta_{\text{link}}$ . The dashed line represents the classical threshold of  $F = 0.5$ , above which a state is entangled. **c**, Maximum fidelity of deterministically delivered states as a function of  $\eta_{\text{link}}$ . A critical threshold of  $\eta_{\text{link}} \approx 0.83$  (vertical green line) must be surpassed to deliver an entangled state at every cycle on average.



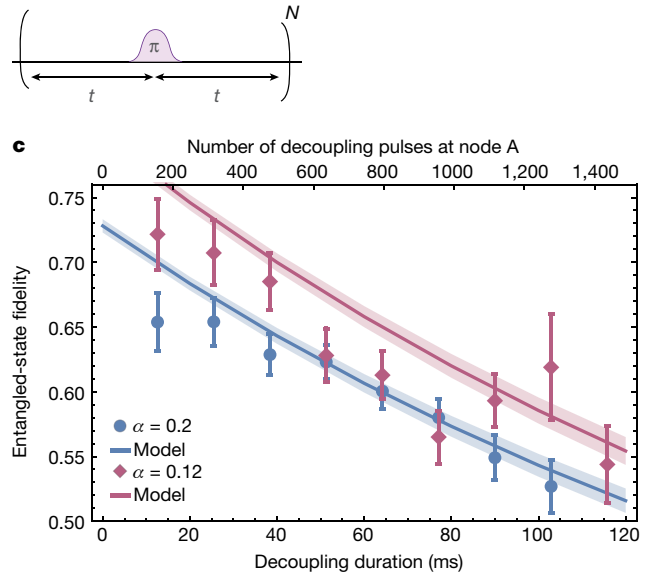
**Fig. 2 | Benchmarking single-photon entanglement generation.** **a**, Experimental protocol. (1) Before entanglement generation, an NV-centre state check verifies that the NV centre is in the correct charge state (the negatively charged state) and resonant with the excitation laser (discussed further in Methods); this state is denoted 'ready'. This is repeated until the check passes. (2) Entanglement generation is attempted until success is heralded, in which case we continue to readout (step (3)). If 250 attempts have been made without success, we revert to step (1). (3) Upon heralded success, the spin states are read out in a chosen basis by using microwaves to rotate the state, followed by single-shot readout (light bulbs indicate the detection of the bright (1) or dark (0) spin state). **b**, The left panel shows the optical set-up used for entanglement generation, in which the optical excitation pulses for each node are derived at a beam splitter. The single photons emitted by the nodes as a result of these excitation pulses are interfered on another beam splitter, completing an

effective interferometer between the nodes. The optical phase difference  $\Delta\theta$  acquired in this interferometer must be known. At pre-determined intervals, light is injected into the interferometer, as shown in the right panel. This light is measured using the same detectors that herald entanglement, and the signal is fed back via a microcontroller to a piezo-electric fibre stretcher that is used to compensate for phase drifts. For the data reported here, we stabilize the phase difference every 180 ms. **c**, Measured  $\langle XX \rangle$  and  $\langle YY \rangle$  correlations (left) for  $\psi_{0/1}$  (where 0/1 denotes the heralding detector) and  $\alpha = 0.1$  as the readout basis is swept at node A (inset). The right panel shows the measured  $\langle ZZ \rangle$  correlations. **d**, **e**, Fidelity of the heralded states with respect to a Bell state (**d**) and entanglement-generation success rate (**e**), for different values of  $\alpha$ . For **c–e**, solid lines (with shaded 1-s.d. statistical uncertainties) are the predictions of our model based solely on independently determined parameters (Methods). Error bars in **c** and **d** represent 1 s.d.





**Fig. 3 | Coherence protection of remote-entangled states.** **a**, Dynamical decoupling protects the state of the NV-centre spins from quasi-static environmental noise. To protect a spin for a time  $T$ ,  $N = T/(2t)$  inverting ( $\pi$ ) microwave pulses are applied at  $2t$  intervals. For node A,  $t = 40.320 \mu\text{s}$ ; for node B,  $t = 36.148 \mu\text{s}$ . **b**, Fidelity with respect to the initial state for dynamical decoupling of the state  $(|\uparrow\rangle + |\downarrow\rangle)/\sqrt{2}$  at each of our nodes. Solid lines show exponential fits with coherence times of 290(20) ms and



680(70) ms for nodes A and B, respectively. **c**, Dynamical decoupling of entangled states created using the single-photon entanglement protocol for  $\alpha = 0.12$  and  $\alpha = 0.2$ . Solid lines (with shaded 1-s.d. statistical uncertainties) show the predictions of our model (Methods) based on the data in **b**, from which the entangled-state coherence time is expected to be  $\tau = 200(10)$  ms. Error bars in **b** and **c** represent 1 s.d.

Single-photon entanglement generation (Fig. 2a) proceeds by first initializing each node in  $|\uparrow\rangle$  by optical pumping<sup>25</sup>, followed by coherent rotation using a microwave pulse<sup>26</sup> to create the state

$$|NV\rangle = \sqrt{\alpha}|\uparrow\rangle + \sqrt{1-\alpha}|\downarrow\rangle \quad (3)$$

where  $\alpha$  is determined by the choice of microwave pulse. We then apply resonant laser light to excite selectively the ‘bright’ state  $|\uparrow\rangle$  to an excited state, which rapidly decays radiatively back to the ground state by emitting a single photon. This entangles the spin state of the NV centre with the presence ( $|\uparrow\rangle$ ) or absence ( $|\downarrow\rangle$ ) of a photon in the emitted optical mode:

$$|NV, \text{optical mode}\rangle = \sqrt{\alpha}|\uparrow\rangle|1\rangle + \sqrt{1-\alpha}|\downarrow\rangle|0\rangle \quad (4)$$

Emitted photons are transmitted to a central station at which a beam splitter is used to remove their ‘which path’ information. Successful detection of a photon at this station indicates that at least one of the NV centres is in the bright state  $|\uparrow\rangle$  and therefore heralds the creation of a spin–spin entangled state. However, given the detection of one photon, the conditional probability that the other NV centre is also in the  $|\uparrow\rangle$  state, but that the photon it emitted was lost, is  $p = \alpha$  (in the limit  $p_{\text{det}} \ll 1$ , where  $p_{\text{det}}$  is the photon detection efficiency). This degrades the heralded state from a maximally entangled Bell state  $|\psi\rangle = (|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle)/\sqrt{2}$  to

$$\rho_{NV,NV} = (1-\alpha)|\psi\rangle\langle\psi| + \alpha|\uparrow\uparrow\rangle\langle\uparrow\uparrow| \quad (5)$$

The probability of successfully heralding entanglement is  $2p_{\text{det}}\alpha$ . The state fidelity  $F = 1 - \alpha$  can therefore be traded off against the entanglement rate directly. The corresponding success probability of a two-photon protocol is  $p_{\text{det}}^2/2$ ; for a given acceptable infidelity  $\alpha$ , single-photon protocols will therefore provide a rate increase of  $4\alpha/p_{\text{det}}$ . For example, for our system’s  $p_{\text{det}} \approx 4 \times 10^{-4}$ , if a 10% infidelity is acceptable, then the rate can be increased by three orders of magnitude compared to two-photon protocols.

The primary challenge in implementing single-photon entanglement is that the resulting entangled state depends on the optical phase acquired by the laser pulses used to create spin–photon entanglement

at each node, as well as on the phase acquired by the emitted single photons as they propagate (Fig. 2b). The experimental set-up therefore acts as an interferometer from the point at which the optical pulses are split to the point at which the emitted optical modes interfere. For a total optical phase difference of  $\Delta\theta$ , the entangled state created is

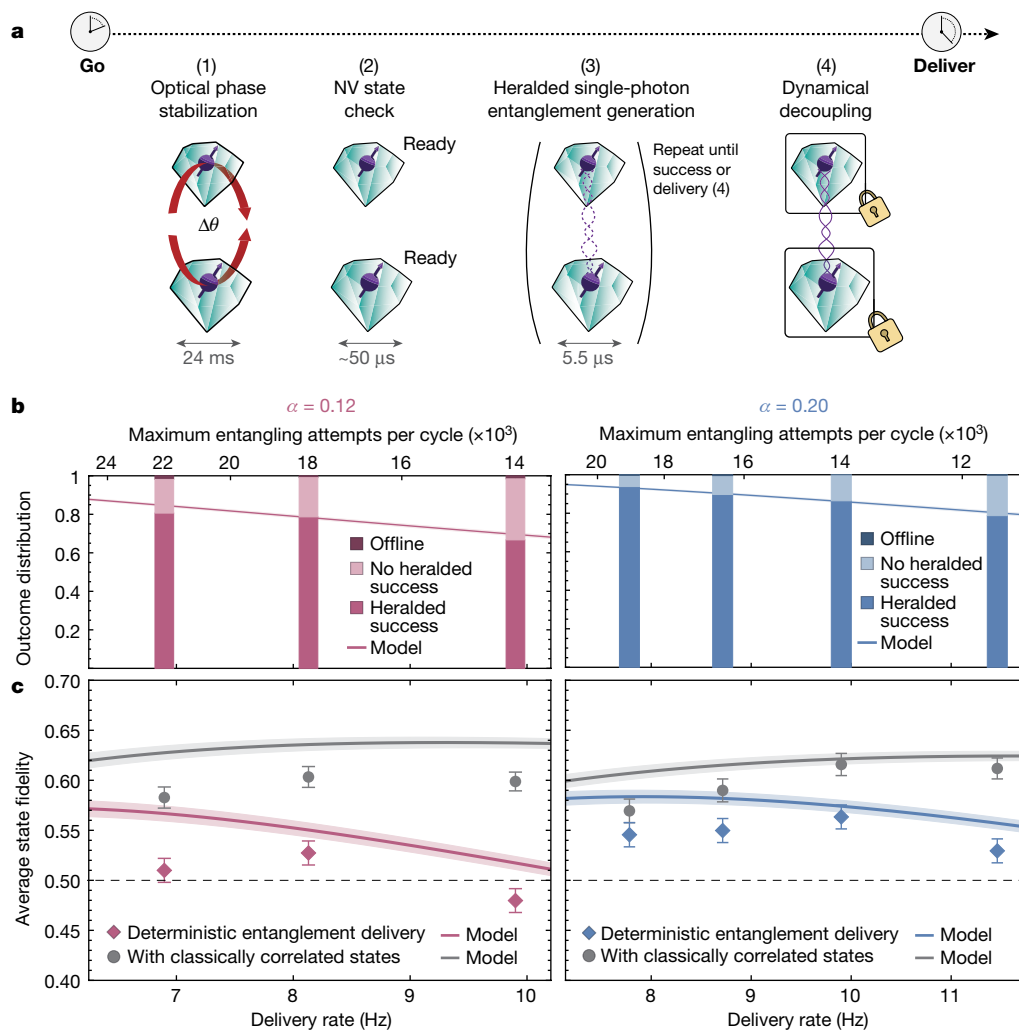
$$|\psi_{0/1}(\Delta\theta)\rangle = |\uparrow\downarrow\rangle \pm e^{i\Delta\theta}|\downarrow\uparrow\rangle \quad (6)$$

where 0/1 (with corresponding  $\pm$  phase factor) denotes the detector at the central station that detected the incident photon. This optical phase difference must be known to ensure that entangled states are available for further use.

We overcome this entangled-state phase sensitivity by interleaving periods of optical-phase stabilization with our entanglement generation. During phase stabilization we input bright laser light at the same frequency as the NV-centre excitation light and detect the light reflected from the diamond substrate using the same detectors that are used to herald entanglement. The measured optical phase, estimated from the detected counts, is used to adjust the phase back to our desired value using a piezoelectric fibre stretcher. We achieve an average steady-state phase stability of  $14.3(3)^\circ$ , limited by the mechanical oscillations of the optical elements in our experimental set-up (the error quoted here and elsewhere is one standard deviation; Methods, Extended Data Fig. 6).

To demonstrate the controlled generation of entangled states, we run the single-photon entangling protocol with a bright-state population of  $\alpha = 0.1$ . After entanglement is heralded, we apply basis rotations and single-shot state readout<sup>25</sup> at each node (A and B) to measure  $\langle\sigma_i^A\sigma_j^B\rangle$  correlations between the nodes, where hereafter the standard Pauli matrices are referred to in the shorthand  $\sigma_X, \sigma_Y, \sigma_Z = X, Y, Z$ . We observe strong correlations for  $\langle XX\rangle$  and  $\langle YY\rangle$  and, when sweeping the readout basis for node A, oscillations of these coherences, as expected from the desired entangled state (Fig. 2c, left). In combination with the measured  $\langle ZZ\rangle$  correlations (Fig. 2c, right), this finding unambiguously demonstrates the establishment of entanglement between our nodes.

We explore the trade-off between the entangled-state fidelity and the entanglement rate by measuring  $\langle XX\rangle$ ,  $\langle YY\rangle$  and  $\langle ZZ\rangle$  correlations for a range of different initial bright-state populations  $\alpha$ . Using these



**Fig. 4 | Deterministic entanglement delivery.** **a**, Each deterministic entanglement-delivery cycle combines: (1) optical phase stabilization; (2) NV-centre state checks, repeated until a threshold number of photons are detected at each node; (3) attempts at probabilistic entanglement generation (Fig. 2); and (4) upon heralded entanglement success, state protection by dynamical decoupling until the delivery time. **b**, Distribution of deterministic entanglement delivery outcomes for  $\alpha = 0.12$  (left) and  $\alpha = 0.2$  (right) and different delivery rates. The fraction of cycles in which a herald photon is detected ('heralded success'), in which no herald is detected ('no heralded success') and in which the NV-

centre state checks for at least one of the NV centres fail repeatedly for the whole cycle ('offline'; often too small to be visible in the plot) are shown. The lines show the success rates predicted by our model. **c**, Average fidelity of deterministically delivered entangled states for  $\alpha = 0.12$  (left) and  $\alpha = 0.2$  (right) and different delivery rates (diamonds). The average fidelity if classically correlated states were delivered for cycles in which no success event was heralded is also shown (circles). The associated lines (with shaded 1-s.d. statistical uncertainties) are the corresponding predictions of our model (Methods). Error bars in **b** and **c** represent 1 s.d.

correlations, we calculate the fidelity of the heralded state relative to the desired maximally entangled Bell state for each value of  $\alpha$  (Fig. 2d), along with the measured success rate (Fig. 2e). As predicted, the fidelity increases with decreasing  $\alpha$  as the weight of the unentangled state  $|\uparrow\uparrow\rangle\langle\uparrow\uparrow|$  diminishes (equation (5)). For small  $\alpha$ , the fidelity saturates because the dark-count rates of the detectors become comparable to the detection rate.

Choosing  $\alpha$  to maximize fidelity, we find that our protocol allows us to generate entanglement with a fidelity of 0.81(2) at a rate of  $r_{\text{ent}} = 6$  Hz (for  $\alpha = 0.05$ ). Alternatively, by trading the entanglement fidelity for rate, we can generate entanglement at  $r_{\text{ent}} = 39$  Hz with an associated fidelity of 0.60(2) ( $\alpha = 0.3$ ). This represents an increase in the entangling rate of two orders of magnitude compared to previous NV-centre experiments<sup>10</sup> and of three orders of magnitude compared to two-photon protocols under the same conditions<sup>14</sup>. Compared to the maximum theoretical fidelity for  $\alpha = 0.05$  of 0.95, the states we generate have a 3% reduction in fidelity due to residual photon distinguishability, 4% from double excitation, 3% from detector dark counts and 2% from optical-phase uncertainty (Methods).

To reach a sufficient link efficiency  $\eta_{\text{link}}$  to enable deterministic entanglement delivery, the single-photon protocol must be combined with robust protection of the remote-entangled states that are generated. To achieve this, we carefully shielded our NV centres from external noise sources, including residual laser light and microwave amplifier noise, leaving as the dominant noise the slowly fluctuating magnetic field induced by the surrounding nuclear spin bath.

We mitigate this quasi-static noise by implementing dynamical decoupling with 'XY8' pulse sequences (Fig. 3a, Methods, Extended Data Fig. 9). The fixed delay between microwave pulses in these sequences is optimized for each node<sup>27</sup>. Varying the number of decoupling pulses allows us to protect the spins for different durations. This dynamical decoupling extends the coherence time of node A and node B from about 5  $\mu\text{s}$  to 290(20) ms and 680(70) ms, respectively (Fig. 3b). The difference in coherence times for the two nodes is attributed to differing nuclear-spin environments and microwave-pulse fidelities.

To investigate the preservation of remote-entangled states, we incorporate dynamical decoupling for varying durations after successful single-photon entanglement generation (Fig. 3c). We find

an entangled-state coherence time of 200(10) ms (decoherence rate  $r_{\text{dec}} = 5.0(3)$  Hz). The observed entangled-state fidelities closely match the predictions of our model, which is based solely on independently determined parameters (Methods, Extended Data Table 1). In particular, the decoherence of the remote-entangled state is fully explained by the combination of the individual decoherence rates of the individual nodes.

The combination of dynamical decoupling and the single-photon entanglement protocol achieves a quantum link efficiency of  $\eta_{\text{link}} \approx 8$ , well above the critical threshold of  $\eta_{\text{link}} \approx 0.83$  and comparable to the published<sup>8</sup> state-of-the-art in ion traps,  $\eta_{\text{link}} \approx 5$ .

These innovations enable the design of a deterministic entanglement-delivery protocol that guarantees the delivery of entangled states at specified intervals, without any post-selection of results or pre-selection based on the nodes being in the appropriate conditions (Fig. 4a). Phase stabilization occurs at the start of each cycle, after which there is a pre-set period before an entangled state must be delivered. This window must therefore include all NV-centre state checks (necessary to mitigate spectral diffusion via feedback control and to verify the charge-state and resonance conditions<sup>9</sup>), entanglement-generation attempts and dynamical decoupling necessary to deliver an entangled state. Fast conditional logic is used to adapt the experimental sequence dynamically on the basis of the detection of a heralding signal<sup>9,10,28,29</sup>. Further details on the experimental implementation are given in Methods and Extended Data Fig. 1.

We run our deterministic entanglement-delivery protocol at two values of  $\alpha$  (0.2 and 0.12) and for delivery rates of 7–12 Hz. We divide the experiment into runs of 1,500 cycles (that is, 1,500 deterministic-state deliveries), for a total dataset of 42,000 cycles.

We first confirm that heralded entanglement occurs with the expected probabilities (Fig. 4b) by determining the fraction of cycles in which entanglement is heralded, in which no entangling attempts succeed and in which entanglement attempts do not occur at all because the NV-centre state check never succeeds. To establish reliable and useful quantum networks, it is important that entangled states can be delivered with high confidence over long periods. The nodes must therefore not be offline, for example, owing to uncompensated drifts in the resonant frequencies of the optical transitions. We therefore do not stop the experiment from running once it starts and include any such offline cycles in our datasets. Their negligible contribution (0.8% of cycles) confirms the robustness of our experimental platform and the effectiveness of our NV-centre frequency and charge-state control (discussed further in Methods).

For each value of  $\alpha$  and for each pre-set delivery interval, we determine the average fidelity of the deterministically delivered states by measuring their  $\langle XX \rangle$ ,  $\langle YY \rangle$  and  $\langle ZZ \rangle$  correlations (Fig. 4c). We find that for  $\alpha = 0.2$  and a rate of 9.9 Hz, we are able to create states with a fidelity of 0.56(1), demonstrating successful deterministic-entanglement delivery.

Our model (solid lines in Fig. 4c) captures the trends of the deterministic entanglement-delivery data effectively. However, the observed state fidelities are slightly lower than the predicted ones, hinting at sources of decoherence that are not included in our model (Methods, Extended Data Fig. 4). Identifying these potential sources will be the subject of future work.

During cycles in which entanglement is not successfully heralded, the spin states are nonetheless delivered and read out. In these cases, we deliver the state that the NV centres are left in after a failed entanglement attempt, which has a low fidelity with respect to the desired Bell state (for example,  $F_{\text{unent}} = 0.04$  for  $\alpha = 0.2$ ). Although this stringent test highlights the robust nature of our protocol, we could instead deliver a mixed state ( $F_{\text{unent}} = 1/4$ ) or a classically correlated state ( $F_{\text{unent}} = 1/2$ ) when a successful event is not heralded. The resulting fidelities for our experimental data if classically correlated states were delivered are also plotted in Fig. 4b (grey circles). In this case, we would be able to deliver entangled states deterministically with fidelities of 0.62(1) at a rate of 9.9 Hz.

The deterministic entanglement delivery between remote NV centres demonstrated here is enabled by a quantum link efficiency exceeding unity. Straightforward modifications to our experiment are expected to increase the quantum link efficiency further. Refinements to the classical experimental control will allow us to reduce the duration of the entanglement attempt from 5.5  $\mu\text{s}$  to less than 2  $\mu\text{s}$ , which would more than double the entangling rate. Furthermore, the entangled-state coherence time could be improved substantially by exploiting long-lived nuclear-spin quantum memories<sup>10,30,31</sup>. We anticipate that this will allow for link efficiencies in excess of 100 in the near future. Further improvements to the photon detection efficiency (including enhancement of the zero-phonon line emission)<sup>32,33</sup> would lead to an additional increase of at least an order of magnitude.

In combination with recent progress on robust storage of quantum states during remote entangling operations<sup>10,34</sup>, the techniques reported here reveal a direct path to the creation of many-body quantum states distributed over multiple quantum network nodes. Moreover, given the demonstrated potential for phase stabilization in optical fibre over distances of tens of kilometres<sup>22</sup>, our results open up the prospect of entanglement-based quantum networks at metropolitan scales.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0200-5>.

Received: 27 April 2018; Accepted: 9 May 2018;

Published online 13 June 2018.

- Kimble, H. J. The quantum internet. *Nature* **453**, 1023–1030 (2008).
- Broadbent, A., Fitzsimons, J. & Kashefi, E. Universal blind quantum computation. In *50th Annual IEEE Symposium on Foundations of Computer Science* 517–526 (IEEE, 2009).
- Jiang, L. et al. Quantum repeater with encoding. *Phys. Rev. A* **79**, 032325 (2009).
- Ekert, A. & Renner, R. The ultimate physical limits of privacy. *Nature* **507**, 443–447 (2014).
- Gottesman, D., Jennewein, T. & Croke, S. Longer-baseline telescopes using quantum repeaters. *Phys. Rev. Lett.* **109**, 070503 (2012).
- Nickerson, N. H., Fitzsimons, J. F. & Benjamin, S. C. Freely scalable quantum technologies using cells of 5-to-50 qubits with very lossy and noisy photonic links. *Phys. Rev. X* **4**, 041041 (2014).
- Kómár, P. et al. A quantum network of clocks. *Nat. Phys.* **10**, 582–587 (2014).
- Hucul, D. et al. Modular entanglement of atomic qubits using photons and phonons. *Nat. Phys.* **11**, 37–42 (2015).
- Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
- Kalb, N. et al. Entanglement distillation between solid-state quantum network nodes. *Science* **356**, 928–932 (2017).
- Reiserer, A. & Rempe, G. Cavity-based quantum networks with single atoms and optical photons. *Rev. Mod. Phys.* **87**, 1379–1418 (2015).
- Hofmann, J. et al. Heralded entanglement between widely separated atoms. *Science* **337**, 72–75 (2012).
- Northup, T. & Blatt, R. Quantum information transfer using photons. *Nat. Photon.* **8**, 356–363 (2014).
- Pfaff, W. et al. Unconditional quantum teleportation between distant solid-state quantum bits. *Science* **345**, 532–535 (2014).
- Monroe, C. et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89**, 022317 (2014).
- Pant, M. et al. Routing entanglement in the quantum internet. Preprint at <https://arxiv.org/abs/1708.07142> (2017).
- Schoute, E., Mancinska, L., Islam, T., Kerenidis, I. & Wehner, S. Shortcuts to quantum network routing. Preprint at <https://arxiv.org/abs/1610.05238> (2016).
- Stockill, R. et al. Phase-tuned entangled state generation between distant spin qubits. *Phys. Rev. Lett.* **119**, 010503 (2017).
- Delteil, A. et al. Generation of heralded entanglement between distant hole spins. *Nat. Phys.* **12**, 218–223 (2016).
- Bar-Gill, N., Pham, L. M., Jarmola, A., Budker, D. & Walsworth, R. L. Solid-state electronic spin coherence time approaching one second. *Nat. Commun.* **4**, 1743 (2013).
- Cabrillo, C., Cirac, J. I., Garca-Fernández, P. & Zoller, P. Creation of entangled states of distant atoms by interference. *Phys. Rev. A* **59**, 1025–1033 (1999).
- Minář, J., de Riedmatten, H., Simon, C., Zbinden, H. & Gisin, N. Phase-noise measurements in long-fiber interferometers for quantum-repeater applications. *Phys. Rev. A* **77**, 052325 (2008).
- Casabone, B. et al. Heralded entanglement of two ions in an optical cavity. *Phys. Rev. Lett.* **111**, 100505 (2013).



24. Sipahigil, A. et al. An integrated diamond nanophotonics platform for quantum-optical networks. *Science* **354**, 847–850 (2016).
25. Robledo, L. et al. High-fidelity projective read-out of a solid-state spin quantum register. *Nature* **477**, 574–578 (2011).
26. Fuchs, G. D., Dobrovitski, V. V., Toyli, D. M., Heremans, F. J. & Awschalom, D. D. Gigahertz dynamics of a strongly driven single quantum spin. *Science* **326**, 1520–1522 (2009).
27. Abobeih, M. H. et al. One-second coherence for a single electron spin coupled to a multi-qubit nuclear-spin environment. Preprint at <https://arxiv.org/abs/1801.01196> (2018).
28. Chou, C.-W. et al. Functional quantum nodes for entanglement distribution over scalable quantum networks. *Science* **316**, 1316–1320 (2007).
29. Matsukevich, D. N. et al. Deterministic single photons via conditional quantum evolution. *Phys. Rev. Lett.* **97**, 013601 (2006).
30. Maurer, P. C. et al. Room-temperature quantum bit memory exceeding one second. *Science* **336**, 1283–1286 (2012).
31. Yang, S. et al. High-fidelity transfer and storage of photon states in a single nuclear spin. *Nat. Photon.* **10**, 507–511 (2016).
32. Riedel, D. et al. Deterministic enhancement of coherent photon generation from a nitrogen-vacancy center in ultrapure diamond. *Phys. Rev. X* **7**, 031040 (2017).
33. Wan, N. H. et al. Efficient extraction of light from a nitrogen-vacancy center in a diamond parabolic reflector. Preprint at <https://arxiv.org/abs/1711.01704> (2017).
34. Reiserer, A. et al. Robust quantum-network memory using decoherence-protected subspaces of nuclear spins. *Phys. Rev. X* **6**, 021040 (2016).

**Acknowledgements** We thank S. van Dam, M. Abobeih, T. Taminiau, F. Rozpedek, K. Goodenough and S. Wehner for discussions. We acknowledge support from the Netherlands Organisation for Scientific Research (NWO) through a VICI grant and from the European Research Council through a Starting Grant and a Synergy Grant.

**Reviewer information** *Nature* thanks D. Englund, J. Laurat and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** P.C.H., N.K. and J.P.J.M. prepared the experimental apparatus. P.C.H. and N.K. carried out the experiments. R.F.L.V. and R.N.S. conceived the microwave switch circuit. P.C.H. analysed the data and wrote the manuscript with input from N.K. and R.H. The diamond substrates were grown by D.J.T. and M.M. The project was supervised by R.H.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0200-5>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to R.H.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

**Deterministically delivered entangled-state fidelity as a function of quantum link efficiency.** We assume an entanglement-generation rate  $r_{\text{ent}}$  and an entangled-state decoherence rate  $r_{\text{dec}}$ . If the rate at which entanglement attempts occur is much faster than  $r_{\text{ent}}$  (that is, there is a low probability of success), then we can approximate entanglement generation as a continuous process. In this case, the probability density for successfully generating entanglement at a time  $t$  after beginning our attempts is  $p_{\text{ent}}(t) = r_{\text{ent}} e^{-r_{\text{ent}} t}$ . The corresponding cumulative probability of success is  $p_{\text{succ}}(t) = 1 - e^{-r_{\text{ent}} t}$ .

Once we succeed at creating entanglement, the state will decohere until the time at which we deliver it. For single-qubit depolarizing noise at each site, the fidelity of the resulting state after storage for a time  $t$  is

$$F(t) = \frac{1}{4} + \frac{3}{4} e^{-r_{\text{dec}} t}$$

If we deliver our entangled state at time  $t_{\text{ent}} = \beta/r_{\text{dec}}$  (where  $\beta$  parameterizes the time in terms of the decoherence rate), then the average fidelity of the delivered state (given a success occurred) is

$$\begin{aligned} F_{\text{succ}} &= \frac{1}{p_{\text{succ}}(t_{\text{ent}})} \int_0^{t_{\text{ent}}} p_{\text{ent}}(t) F(t_{\text{ent}} - t) dt \\ &= \frac{1}{p_{\text{succ}}(t_{\text{ent}})} \int_0^{t_{\text{ent}}} r_{\text{ent}} e^{-r_{\text{ent}} t} \left[ \frac{1}{4} + \frac{3}{4} e^{-r_{\text{dec}}(t_{\text{ent}} - t)} \right] dt \\ &= \frac{3e^{-\beta} \eta_{\text{link}} + (1 - 4\eta_{\text{link}}) e^{-\eta_{\text{link}} \beta} + \eta_{\text{link}} - 1}{4(\eta_{\text{link}} - 1) p_{\text{succ}}(t_{\text{ent}})} \end{aligned}$$

Because  $p_{\text{succ}}(t_{\text{ent}}) = 1 - e^{-\beta \eta_{\text{link}}}$ ,  $\beta = -\ln[1 - p_{\text{succ}}(t_{\text{ent}})]/\eta_{\text{link}}$ . Using this, along with the shorthand  $p_{\text{succ}} = p_{\text{succ}}(t_{\text{ent}})$ , we find that

$$F_{\text{succ}} = \frac{3\eta_{\text{link}} + p_{\text{succ}} - 3\eta_{\text{link}}(1 - p_{\text{succ}})^{1/\eta_{\text{link}}} - 4\eta_{\text{link}} p_{\text{succ}}}{4p_{\text{succ}}(1 - \eta_{\text{link}})}$$

As discussed in the main text, we can choose to draw a black box around this process, delivering an unentangled state (state fidelity  $F_{\text{unent}} \leq 1/2$ ) for cycles in which no attempt at entanglement-generation succeeds so that a state is always delivered. This means that the states output from this black box will have a fidelity with respect to a Bell state  $F_{\text{det}}$  given by equation (1), where  $F_{\text{succ}}$  is as in the above equation. The maximum achievable fidelity when outputting a fully mixed state ( $F_{\text{unent}} = 1/4$ ) upon failure  $F_{\text{det}}^{\text{max}}$  (equation (2)) is found by optimizing  $F_{\text{succ}}$  for a given quantum link efficiency  $\eta_{\text{link}}$ .

The full state of a quantum system can only be experimentally determined using an ensemble of identical states. This means that, in the absence of information about which deterministic entanglement-delivery cycles have a heralded success, the only accurate description of the output of such a black-box system is that a statistical mixture is deterministically output at each cycle.

**Experiment design.** We use chemical-vapour-deposition homoepitaxially grown diamonds of type IIa with a natural abundance of carbon isotopes. Both diamonds were cut along the  $\langle 111 \rangle$  crystal axis and were grown by Element Six. They are situated in custom-built confocal microscope set-ups within closed-cycle cryostats (4 K, Montana Instruments) separated by 2 m. We use fast microwave switches to shield both NV centres from microwave amplifier noise and therefore increase the coherence times substantially (node A uses Qorvo TGS2355-SM and node B uses Analogue Devices HMC544). All other parts of the set-up and sample details are described in the supplementary information of refs <sup>9,10</sup>.

One cycle of the deterministic entanglement protocol consists of optical phase stabilization (described further below), charge-resonance checks to ensure that both NV centres are in the appropriate charge state and on-resonance<sup>25</sup>, heralded single-photon entanglement generation and finally dynamical decoupling to protect the state of the NV centres from their environment until the delivery time. The experimental sequences used in each step of this protocol (and the single-photon entanglement-generation experiment) are detailed in Extended Data Fig. 1.

After delivery, the state of each NV centre is measured in a chosen basis. We use spin-selective optical readout of the NV-centre electron spin to determine its state in a single shot via the optical  $E_x$  transition on both nodes<sup>25</sup>. We measure single-shot readout fidelities of 0.959(3) (0.950(3)) for the bright  $|m_s = 0\rangle \equiv |\uparrow\rangle$  ground state and 0.995(1) (0.996(1)) for the dark  $|m_s = -1\rangle \equiv |\downarrow\rangle$  state on node A (node B). These values are subsequently used to correct for readout errors of the electron spins in state-tomography measurements.

**Experiment control and communication logic.** Extended Data Fig. 2 gives the decision trees and control logic for the ADwin microprocessors (Jaeger ADwin Pro II) that control the experiments. These microcontrollers are responsible for

controlling all other experimental hardware and also communicate with each other to synchronize the experiment.

**Herald photon-detection window.** We use a combination of polarization and temporal filtering to separate the excitation pulse from photons emitted by the NV centre. This necessitates a compromise between collecting as much of the emission light as possible, while ensuring that contamination from the pulse is minimized. In our experiment, we choose a temporal filter window (Extended Data Fig. 3) so that the pulse (assumed to have a Gaussian profile) is suppressed to the level of the detector dark counts by the beginning of the window. The end of the window about 30 ns after the pulse is chosen so that, for all of the datasets collected, the rate of detected NV-centre photons is greater than ten times the dark-count rate at all points within the window. We use a complex programmable logic device to apply this temporal filtering during our experiment and herald the successful generation of an entangled state in real time.

**Theoretical model of deterministic entanglement delivery.** We develop a detailed model to determine the expected performance of the deterministic entanglement-delivery experiment, based on the independently measured parameters given in Extended Data Table 1.

Once the set-ups are determined to be ready, the core entanglement sequence begins with single-photon entanglement generation. This proceeds by first initializing each node in  $|\uparrow\rangle$ , followed by a coherent rotation using a microwave pulse to create the state given in equation (3). Resonant excitation of the NV-centre nodes excites only the bright  $|\uparrow\rangle$  level to an excited state, which rapidly decays radiatively back to the ground state by emitting a single photon. This entangles the state of the NV centre with the presence ( $|1\rangle$ ) or absence ( $|0\rangle$ ) of a photon in the emitted optical mode (equation (4)). The photons emitted by each NV centre are transmitted to a central station at which a beam splitter is used to remove their ‘which path’ information. Successful detection of a photon at this station indicates that at least one of the NV centres is in the bright  $|\uparrow\rangle$  state and thus heralds the creation of a spin–spin entangled state. This entangled state, expressed as  $|\text{NV}_{\text{node A}}, \text{NV}_{\text{node B}}\rangle$ , is (in un-normalized form)

$$\rho = |\psi^\pm\rangle\langle\psi^\pm| + p_{\uparrow\uparrow} |\uparrow\uparrow\rangle\langle\uparrow\uparrow| + p_{\downarrow\downarrow} |\downarrow\downarrow\rangle\langle\downarrow\downarrow|$$

where

$$|\psi^\pm\rangle\langle\psi^\pm| = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & p_{\uparrow\downarrow} & \pm\sqrt{V p_{\uparrow\downarrow} p_{\downarrow\uparrow}} & 0 \\ 0 & \pm\sqrt{V p_{\uparrow\downarrow} p_{\downarrow\uparrow}} & p_{\downarrow\uparrow} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

This state is parameterized by

$$\begin{aligned} p_{\uparrow\uparrow} &= \alpha^2 \{ (1 - p_{\text{dc}})^2 [p_{\text{det}}^{\text{A}} (1 - p_{\text{det}}^{\text{B}}) + p_{\text{det}}^{\text{B}} (1 - p_{\text{det}}^{\text{A}})] \\ &\quad + 2(1 - p_{\text{dc}}) p_{\text{dc}} (1 - p_{\text{det}}^{\text{A}})(1 - p_{\text{det}}^{\text{B}}) \} \\ p_{\uparrow\downarrow} &= \alpha(1 - \alpha) [(1 - p_{\text{dc}})^2 p_{\text{det}}^{\text{A}} + 2p_{\text{dc}} (1 - p_{\text{dc}})(1 - p_{\text{det}}^{\text{A}})] \\ p_{\downarrow\uparrow} &= \alpha(1 - \alpha) [(1 - p_{\text{dc}})^2 p_{\text{det}}^{\text{B}} + 2p_{\text{dc}} (1 - p_{\text{dc}})(1 - p_{\text{det}}^{\text{B}})] \\ p_{\downarrow\downarrow} &= 2(1 - \alpha)^2 p_{\text{dc}} (1 - p_{\text{dc}}) \end{aligned}$$

where  $V$  is the visibility of two-photon interference,  $p_{\text{dc}}$  is the dark-count probability per detector (given by the product of the dark-count rate  $\nu_{\text{dark}}$  and the 25-ns length of the detection window), and  $p_{\text{det}}^{\text{A}}$  and  $p_{\text{det}}^{\text{B}}$  are the probabilities of detecting a photon emitted by node A and node B, respectively. In the limit  $p_{\text{det}} \ll 1$ , for balanced detection probabilities  $p_{\text{det}}^{\text{A}} = p_{\text{det}}^{\text{B}} = p_{\text{det}}$  and assuming no other imperfections,  $\rho$  tends to equation (5).

The corresponding probability of successfully heralding entanglement is

$$\begin{aligned} p_{\text{herald}} &= (1 - p_{\text{dc}}) \{ \alpha(p_{\text{det}}^{\text{A}} + p_{\text{det}}^{\text{B}} - 2p_{\text{det}}^{\text{A}} p_{\text{det}}^{\text{B}} \alpha) \\ &\quad + p_{\text{dc}} [2 - 3(p_{\text{det}}^{\text{A}} + p_{\text{det}}^{\text{B}}) \alpha + 4p_{\text{det}}^{\text{A}} p_{\text{det}}^{\text{B}} \alpha^2] \} \end{aligned}$$

The modelled success rate (plotted in Fig. 2e) is calculated by dividing  $p_{\text{herald}}$  by the entangling-attempt duration (5.5  $\mu\text{s}$ ).

We model double excitation (discussed further below) by applying a Pauli  $Z$  transformation to each of the NV-centre states with probability  $p_{2\text{ph}}/2$ . Phase instability is modelled similarly, by applying a Pauli  $Z$  transformation to one of the states with probability

$$\frac{1}{2} \left[ 1 - \exp \left[ \frac{-(\nu_{\text{int}} t_{\text{p}})^2 - \sigma_{\text{int}}^2}{2} \right] \right]$$

where  $t_{\text{p}}$  denotes the time since phase stabilization.

Finally, we model the effect of dynamical decoupling by assuming that it acts as a depolarizing channel for each qubit<sup>27</sup>. We therefore apply single-qubit depolarizing errors with a probability determined by the measured dynamical-decoupling coherence times. For decoupling for a total time of  $t_d$ , the total probability of a depolarizing error (that is, the application of a Pauli  $X$ ,  $Y$  or  $Z$  transformation with an equal probability) is  $3(1 - e^{-t_d/T_2})/4$ .

This model, based only on independently determined parameters (Extended Data Table 1), captures the trends of our deterministic entanglement generation data effectively (Fig. 4). However, we find that its predictions are slightly offset from the experimental measurements, suggesting that it does not include a small source of infidelity that is present in the experimental data. One potential origin of this discrepancy could be the increased number of attempts (up to two orders of magnitude) at generating entanglement after NV-centre state verification made here as compared to previous experiments<sup>9,10</sup>. Any additional sources of infidelity that may occur over this period (for example, owing to the passive charge-state stabilization process, discussed further below) are not included in the model. A detailed study of these potential imperfections is outside the scope of this work. Nonetheless, as an estimate of the order of this effect, we find that a small systematic correction of 3% to the heralded entangled-state fidelity is sufficient to effectively match our model to the data (Extended Data Fig. 4).

**Passive charge-state stabilization of individual NV centres.** The negatively charged NV centre ( $NV^-$ ) can be ionized under optical illumination via a two-photon absorption process<sup>35</sup>. Owing to the different level structure of the neutral charge state  $NV^0$ , the NV centre will remain dark if such an ionization event occurs during one of our entangling attempts. Ionization therefore hampers the performance of our deterministic entangling protocol by diminishing the success rate and delivery of a separable state upon success. Previous experiments<sup>14</sup> with NV centres that worked in the regime of probabilistically generated yet heralded remote entanglement overcame NV-centre ionization by frequent charge-state verification between protocols and by actively converting the NV centre back to  $NV^-$  via interleaved resonant excitation of the optical transitions of  $NV^0$ .

Such active stabilization protocols would require additional logical overhead in our scenario, where entanglement is generated deterministically. Instead, we passively stabilize the charge state during our entangling sequence by shining in an additional weak laser beam that is resonant with the optical transition of  $NV^0$  (Extended Data Fig. 5). This provides negligible disturbance to the spin-initialization fidelity of  $NV^-$  while bringing the NV centre back into  $NV^-$  if it was converted to  $NV^0$ . We additionally identify that the optical reset beam (duration, 1.5  $\mu$ s) is the main cause of ionization in our system and carefully balance the power of both beams so that the spin state is still well initialized and that ionization is a negligible process for our deterministic entangling protocol (up to 15,000 entangling attempts). Reducing the applied power further by elongating the spin-reset duration would decrease the entanglement rate and limit our quantum link efficiency.

Extended Data Fig. 5 depicts the basic element that, in repetition, forms our sequence to probe the ionization rate. We use simultaneous charge- and spin-reset beams followed by a single microwave  $\pi$  rotation that brings the NV centre into  $|\downarrow\rangle$  and thus guarantees optical excitation during the next round. The NV centre is then read out after a final optical reinitialization into the bright state  $|\uparrow\rangle$ . By increasing the number of sequence repetitions, we observe a decay in the final readout fidelity that is associated with the ionization rate. By increasing the optical intensity of the charge-state reset beam, we obtain a negligible decay as a function of sequence repetitions, allowing us to overcome ionization in our deterministic entangling protocol. The illumination strength of the charge-reset beam is weak enough to avoid inducing noticeable spectral diffusion of the NV-centre emission; our measured entangled states are consistent with a high degree of indistinguishability for both NV-centre emission profiles (discussed further below).

**Optical-phase stabilization.** The single-photon entanglement experiment requires that the optical phase of an effective interferometer between the two nodes is known (Fig. 2). The optical-phase difference between the paths of this interferometer must be known to ensure that entangled states are available for further use. This is achieved by interleaving periods of optical-phase stabilization with our entanglement generation.

For phase stabilization we input bright laser light at the same frequency as, but orthogonally polarized to, the light used for excitation of the NV centres. The orthogonal polarization is chosen because we use a crossed polarizer to filter out the excitation light from the NV-centre emission. Using orthogonally polarized light for phase stabilization allows us to collect more light reflected from the diamond substrate. Before doing this, we verified that there is no measurable difference in the relative phase of the two polarizations within our interferometer.

Measurements of the phase drift (Extended Data Fig. 6a) show a slow drift on second timescales, but several strong resonances at hundreds of hertz (Extended Data Fig. 6b). These resonances are thought to be from mechanical elements in the path of the beam, including the microscope-objective mount. As we were unable to completely suppress these resonances in the current set-ups, we need to measure

the phase over a complete oscillation to estimate the mean phase reliably. The phase must therefore be measured for approximately 10 ms.

We calculate an estimate of the phase from the counts detected at the heralding single-photon detectors. This estimate is used to adjust the phase back to our desired value using a custom-built piezoelectric fibre stretcher and a proportional-integral-derivative routine within our ADwin microcontroller. We find that it takes two to three proportional-integral-derivative cycles to stabilize the phase optimally. We stabilize the phase for three cycles during the single-photon entanglement experiment and for two cycles during the deterministic entanglement experiment. This difference is because phase stabilization occurs during every cycle of the deterministic entanglement-delivery experiment (about 100 ms), whereas it occurs only every 180 ms during the single-photon entanglement experiment and so the phase drifts slightly less after one experimental cycle.

We achieve an average steady-state phase stability of  $14.3(3)^\circ$ , as measured by calibration routines spaced throughout the measurement of our dataset (Extended Data Fig. 6c, d). This stability is limited by the previously identified mechanical oscillations of the optical elements in our experimental set-up. The standard deviation of the phase averaged over a 10-ms period during active stabilization is  $4.8(1)^\circ$ .

Optical phase stabilization is also likely to be feasible for long-distance network links. Using long-wavelength off-resonant light for phase measurements would enable continuous stabilization during entanglement attempts with a negligible effect on the NV-centre state. An experimental study<sup>22</sup> has shown that two network nodes separated by 36 km over a commercial fibre network would still allow for interference visibilities of 99%. For longer distances, it would also be possible to track the phase passively at the time of entanglement delivery and feed this information back to the nodes in which the state is stored, requiring only a coherence time longer than the communication time.

**Two-photon quantum interference.** The quality of photon-mediated heralded entanglement between two emitters hinges on the indistinguishability of their emitted photons. We probe this indistinguishability by interfering emitted single photons on a beam splitter and measuring the number of events in which single-photon detectors connected to the output ports of the beam splitter both detect a photon. For completely indistinguishable single photons, Hong–Ou–Mandel interference ensures that both photons always exit from the same port of the beam splitter, so no coincident events should be detected.

Our two-photon quantum interference experiment proceeds by exciting each emitter with a series of well-separated optical excitation pulses (separated by 1  $\mu$ s). We collect statistics on coincidence events in which one detector registers a photon after one excitation pulse and then the other registers a photon after a later excitation pulse. For an infinite pulse train, the number of coincidence events detected for each number of pulses between the detection events should be constant. However, for a finite pulse train, there are some pulses for a given pulse separation for which there is no partner excitation pulse and therefore no coincident events will be detected. This leads to a linearly decreasing number of coincidence events as a function of pulse difference (Extended Data Fig. 7a).

We use a linear fit to the coincidence events to infer the number of coincidences that would be detected from the same pulse (pulse difference of zero) if fully distinguishable single photons were input (Extended Data Fig. 7b). Because these are single photons, a counting argument shows that, for balanced emission probabilities from each emitter, the expected number of events is half of the value of the linear fit at zero pulse difference.

The ratio  $r$  between the measured number of coincident events within the same pulse and the expected number of events for fully distinguishable photons is related to the single-photon wavefunction overlap  $V = |\langle\psi_a|\psi_b\rangle|^2$  by  $V = (1 - r)$  (again for balanced emission probabilities from each emitter). Incorporating the effect of the known imbalance in emission probabilities in our experiment, we find  $V = 0.90(2)$ .

**Dephasing of entangled states due to double excitation.** An optical Rabi pulse is used to excite the NV-centre nodes to a higher level via a spin-conserving transition. The NV centre subsequently decays back to its original level through spontaneous emission, thereby entangling the spin state of the NV centre and the emitted optical mode. For optical Rabi pulses of finite duration, there is a chance that the NV centre will spontaneously emit a photon during the optical pulse and be re-excited before the end of the pulse. The first emitted photon will be lost to the environment, because it is impossible to distinguish it from the excitation light. However, if the subsequent emitted photon is detected in this double-excitation process, this will falsely herald entanglement. We measured the width of our optical pulse (Extended Data Fig. 8) and used a quantum-jump-based simulation to calculate the corresponding double-excitation probability. Given that the NV centre emitted a photon within the detection window, the probability that double excitation occurred is  $p_{2ph} = 0.04$ .

**State storage via dynamical decoupling.** The coherence time of NV centres is limited by interactions with other magnetic impurities. In our samples, the dominant source of magnetic field noise is the surrounding bath of slowly fluctuating  $^{13}\text{C}$



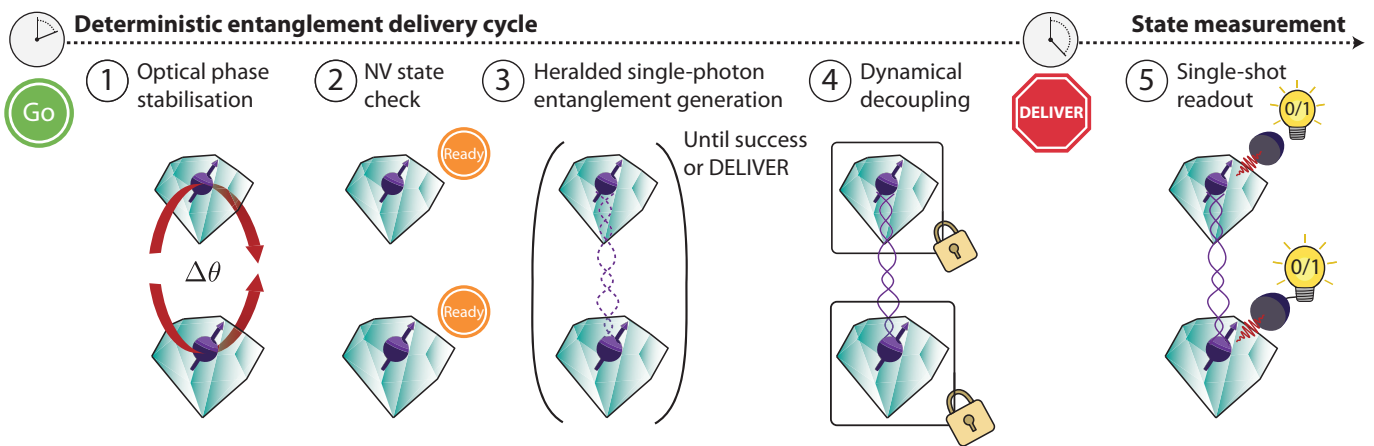
nuclear spins (natural abundance of 1.1%), which results in typical coherence times of 5  $\mu$ s. We use dynamical-decoupling 'XY8' sequences of the form  $(t-\pi_X-2t-\pi_Y-2t-\pi_X-2t-\pi_Y-2t-\pi_X-2t-\pi_Y-2t-\pi_X-2t-\pi_X-t)^{N/8}$  to elongate the coherence times of both NV centres (Fig. 3), with microwave inversion pulses  $\pi$ , the waiting time  $t$  and the number of pulses  $N$ . Each decoupling duration is obtained by arbitrary combinations of  $t$  and  $N$ . We find the optimal combination for a targeted protection duration of about 100 ms by varying  $t$  for a fixed  $N=1,024$ . We choose  $N=1,024$  because the infidelity introduced from inversion-pulse errors is moderate for both nodes.

Extended Data Fig. 9 shows the results of our decoupling-optimization procedure. We prepare the NV centre in a balanced superposition and choose waiting times that are integer multiples of the inverse  $^{13}\text{C}$ -nuclear-spin Larmor frequency  $\nu_L$  to avoid coupling with the nuclear-spin bath (node A,  $\nu_L=443.342$  kHz; node B,  $\nu_L=442.442$  kHz). Following previously reported techniques<sup>27</sup>, we further avoid

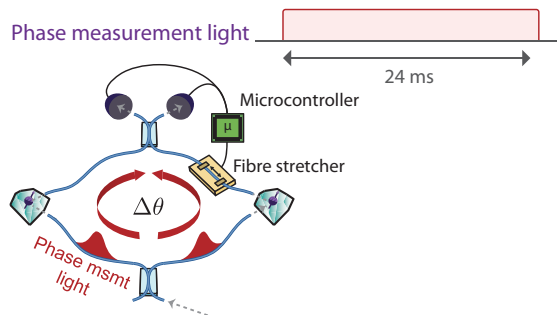
coupling to other magnetic noise sources that result in loss of NV-centre coherence by picking five waiting times with a total variation of 16 ns for each multiple of the inverse Larmor frequency. The data (grey) are then sorted for the waiting time with the best state-preservation quality (blue) at each multiple, giving the minimal NV-centre coherence decay for this number of inversion pulses. We then pick the waiting time that guarantees a low number of inversion pulses while still providing high-quality state protection (red).

**Data availability.** The data sets generated and analysed during this study are available from the corresponding author on reasonable request.

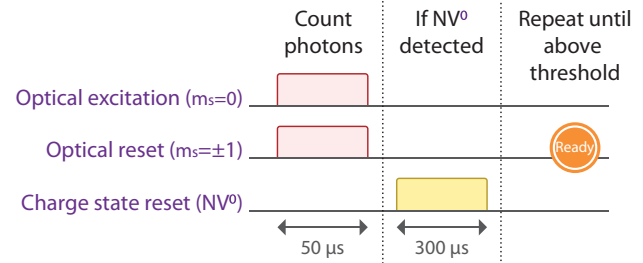
35. Aslam, N., Waldherr, G., Neumann, P., Jelezko, F. & Wrachtrup, J. Photo-induced ionization dynamics of the nitrogen vacancy defect in diamond investigated by single-shot charge state detection. *New J. Phys.* **15**, 013064 (2013).



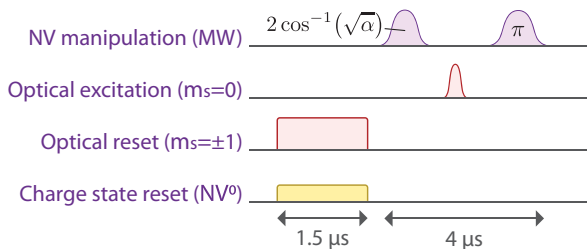
### 1 Optical phase stabilisation



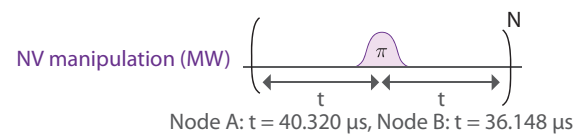
### 2 NV charge/resonance (CR) state check



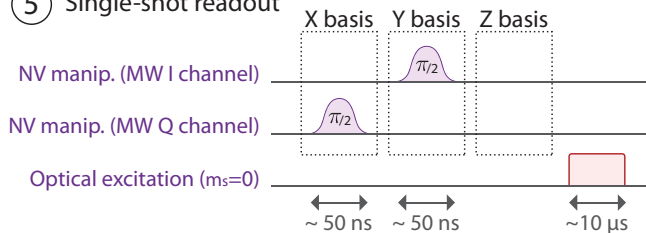
### 3 Heralded single-photon entanglement generation



### 4 State preservation via dynamical decoupling

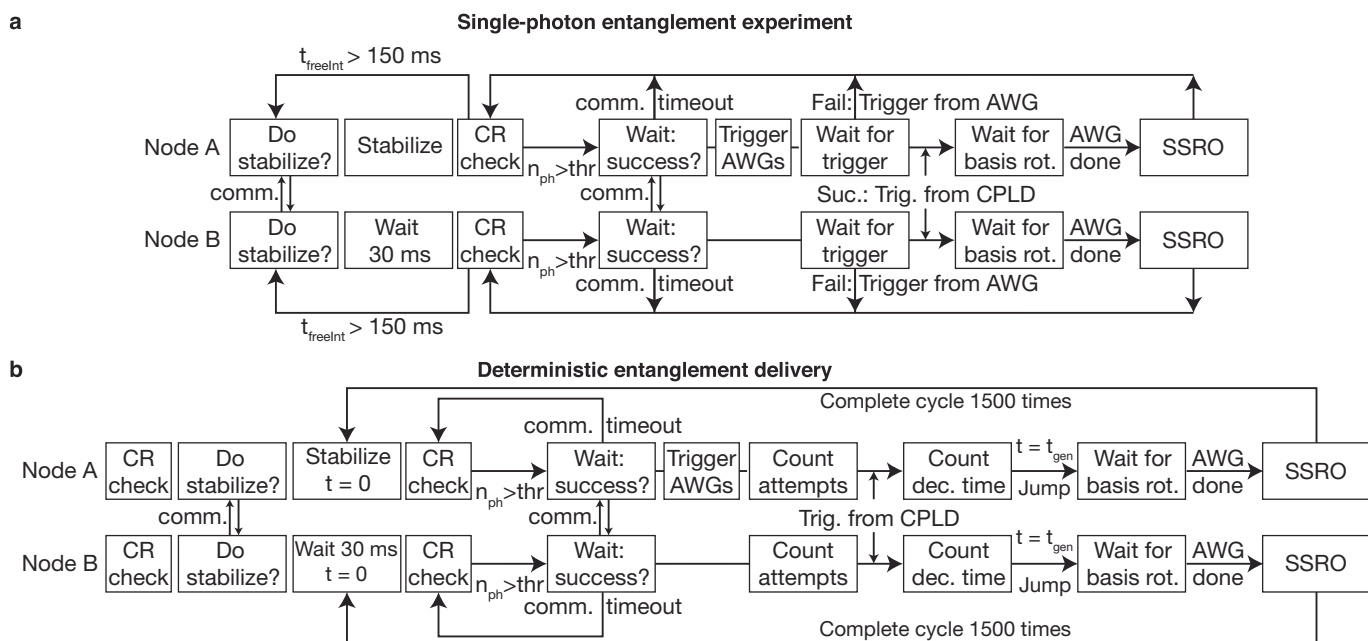


### 5 Single-shot readout



**Extended Data Fig. 1 | Deterministic entanglement-delivery sequences.** Pulse sequences for each step of the deterministic entanglement delivery protocol are shown. These sequences are also used in the single-photon entanglement-generation experiment. (1) Optical phase stabilisation. Bright light is input to measure and stabilize the interferometer (see Methods). The duration is different for the single-photon entanglement experiment. (2) NV-centre state check. By shining in two lasers that are together resonant with transitions from all of the ground states, the NV centre will fluoresce regardless of its ground-state occupation. By counting photons emitted by the NV centre we can verify that both NV centres are in the desired charge state  $NV^-$  and that they are on resonance with the applied lasers. The NV centre is deemed to be on resonance if the number of photons detected during the charge/resonance check surpasses a certain threshold. If no photons are detected, then the

NV centre is assumed to be in the  $NV^0$  state and a resonant laser is applied to reset it to  $NV^-$ . (3) Heralded single-photon entanglement generation. Entanglement generation proceeds by optically re-pumping the spins to  $|\uparrow\rangle$  (including passive charge-state stabilization; see Methods) before a microwave (MW) pulse is used to create the desired bright-state population  $\alpha$  at each node. A resonant excitation pulse then generates spin-photon entanglement. A subsequent microwave  $\pi$  pulse is used to ensure that the NV-centre state is refocused before the next stage should success be heralded. (4) Dynamical decoupling. Microwave pulses are used to implement dynamical decoupling (see Methods). (5) Single-shot readout. The NV-centre nodes can be read out in arbitrary bases in a single shot. If required, a microwave pulse is applied to rotate the qubit state before a resonant laser is applied. Fluorescence photons from the NV centre are detected if it is in the state  $|\uparrow\rangle$ .

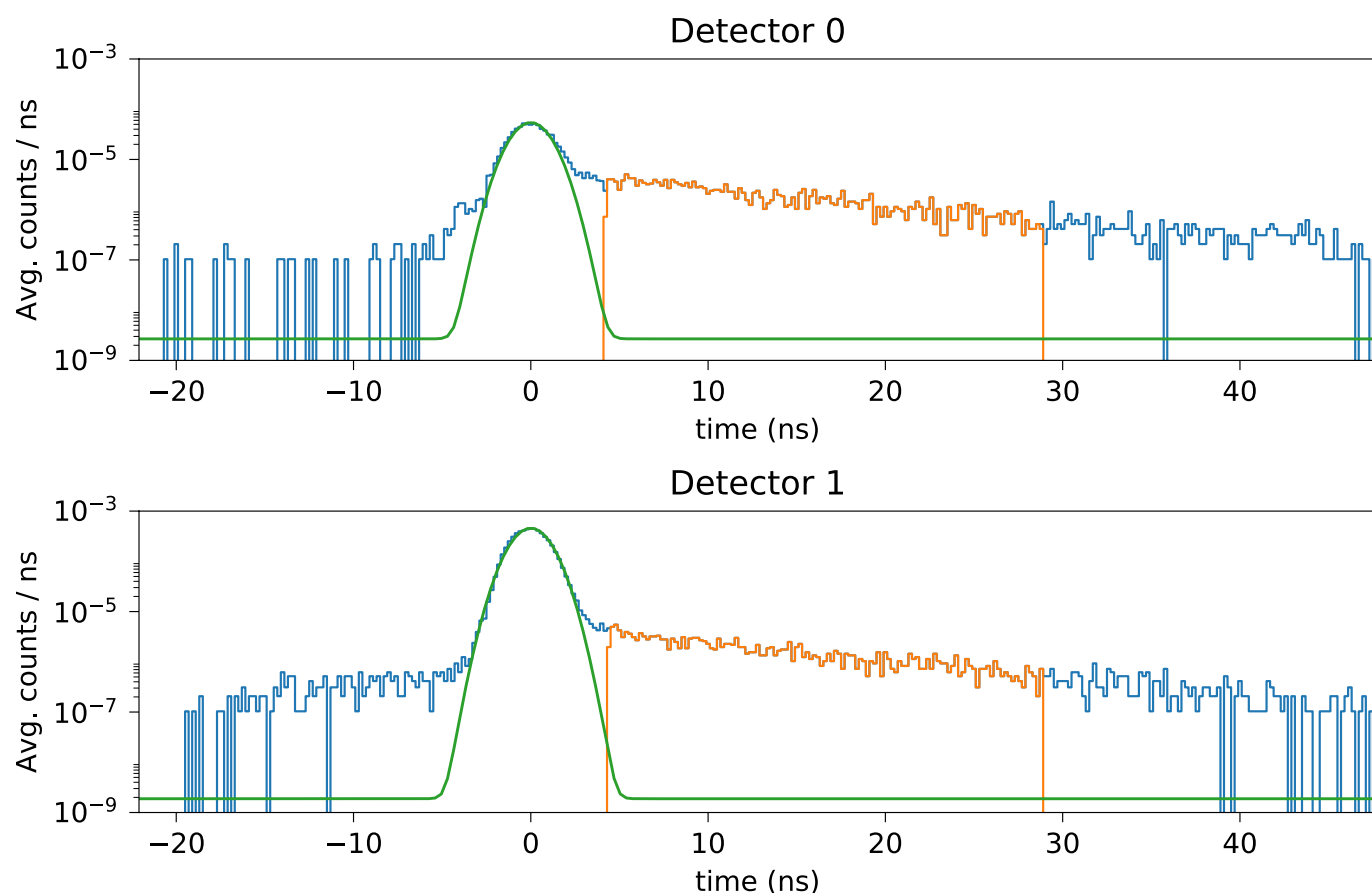


### Extended Data Fig. 2 | Flowchart of the experimental sequences.

The decision trees of the ADwin microprocessors (Jaeger ADwin Pro II) that create the overarching measurement and control loops for network nodes A and B are shown. Both nodes use arbitrary-waveform generators (AWGs) for microwave and laser pulse sequencing (Tektronix AWG5014C). We additionally use a complex programmable logic device (CPLD) to herald the successful generation of an entangled state in real time (described further in Methods). **a**, Decision tree when benchmarking the entangled state. **b**, Deterministic entanglement delivery. Here the ADwin microprocessors keep track of the time since the end of the phase stabilization ( $t = 0$ ). ‘CR check’: as explained in Extended Data Fig. 1, the NV centre is deemed to be on resonance with the excitation lasers if the number of photons detected during the charge/resonance check surpasses a certain threshold ( $n_{ph} > thr$ ); this is repeated until the threshold is passed. ‘comm.’ and ‘comm. timeout’: both ADwin microprocessors exchange classical communication, such as the success of the charge/resonance check, via a three-step-handshake; if one microprocessor waits longer than 1 ms for a response from its counterpart, then the communication times out and we return to the previous logical step (arrow). ‘Count attempts’: the number of entangling attempts  $N$  are counted until  $N = N_{max}$ . ‘Count dec. time’: the time since

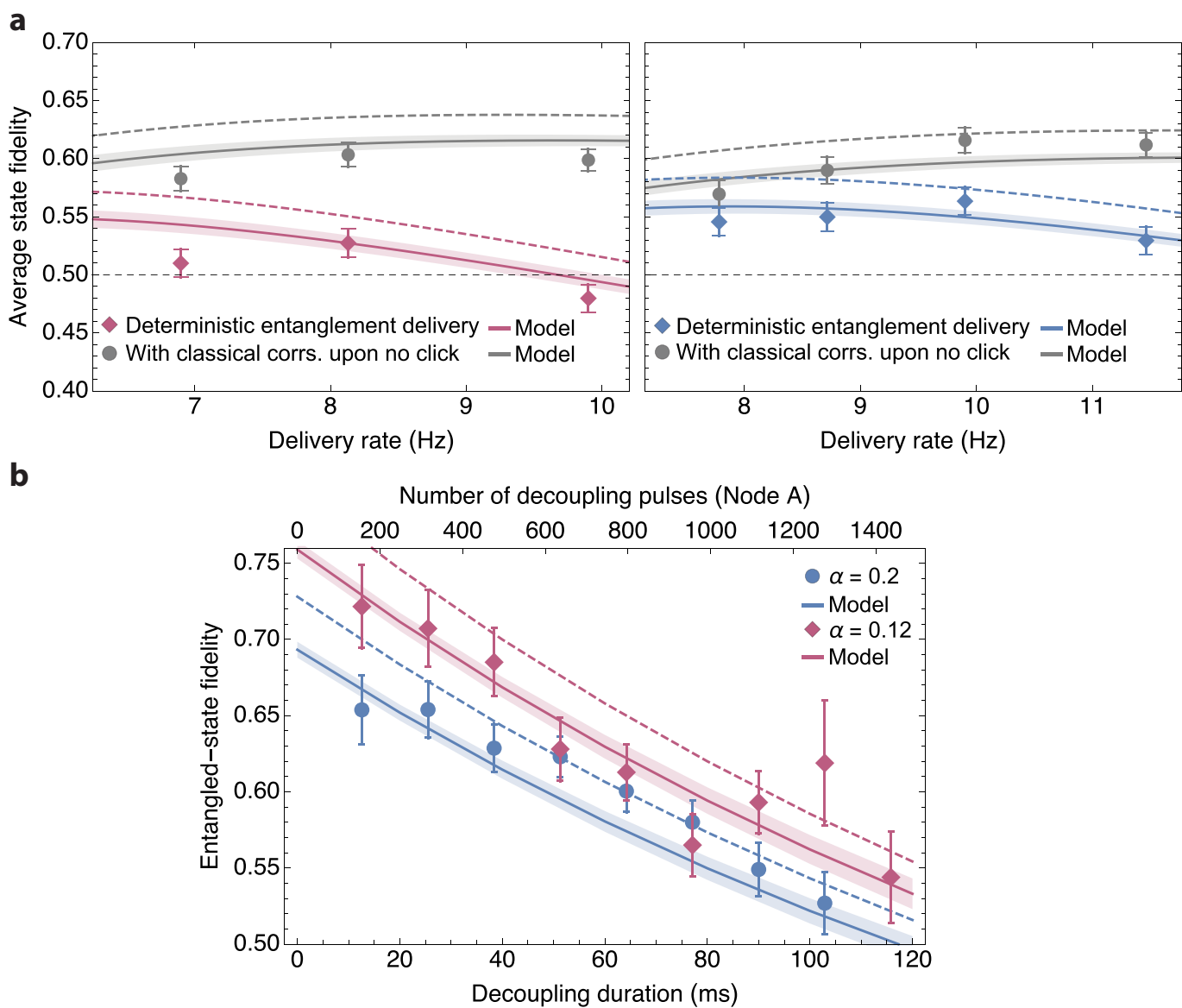
phase stabilization is tracked; if the time is equal to the pre-specified state-generation time  $t_{gen}$ , then the AWG is triggered and the local readout sequences are executed. ‘Wait for basis rot.’: ADwin microprocessors wait for a trigger input from the AWG (‘AWG done’), which indicates that the last microwave rotation before optical readout has been completed. ‘Trigger AWGs’: the microprocessor of node A triggers the AWGs of both nodes to initiate the microwave and entangling sequences; we use a single microprocessor as the trigger source to avoid timing jitter between both generated sequences. ‘SSRO’: optical single-shot readout. ‘Suc.: Trig. from CPLD’ and ‘Fail: Trigger from AWG’: during entanglement generation, the CPLD communicates successful detection of a photon to the nodes; during the single-photon entanglement-benchmarking experiment, the AWG at each node flags failure of the round after 250 entangling attempts. ‘Do stabilize?’: The microprocessors communicate that phase stabilization will be the next step in the experimental sequence; the microprocessor at node A then proceeds with the phase stabilization while that at node B waits until the phase stabilization has finished. The deterministic entangling sequence is run a total of 1,500 times (500 times per readout basis) before a new round is called in, which starts again with the verification of resonant conditions for both NV centres.





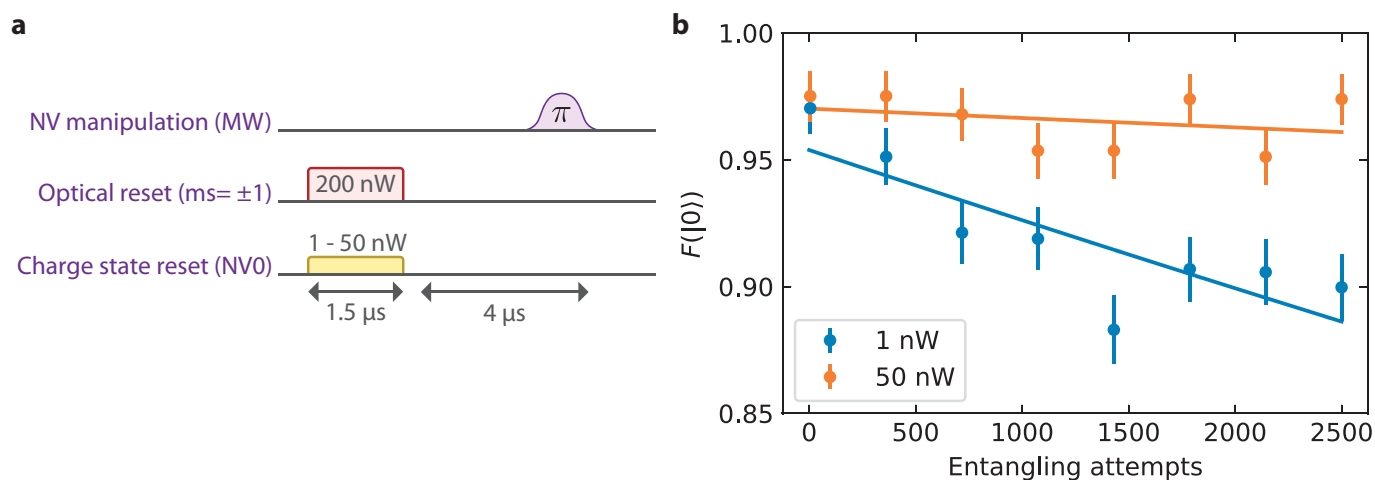
**Extended Data Fig. 3 | Temporal filtering of photons.** Histograms are shown of the times at which photons are detected at each single-photon detector (blue) during a deterministic entanglement-delivery experiment with bright-state population  $\alpha = 0.12$ . The orange histograms show the photons that were detected within the temporal filter window and so were

counted as valid entanglement events. The green line shows a Gaussian fit to the pulse with a full-width at half-maximum of 2.26 ns as measured in Extended Data Fig. 8. This is used to estimate the contribution of residual pulse photons within the filter window.



**Extended Data Fig. 4 | Comparison of experimental model and data.** Including a 3% source of infidelity in our model (which otherwise consists only of independently determined parameters) is sufficient to account for the offset observed between our model and our experimental data. **a**, The modified model, plotted with experimental data reproduced from Fig. 4.

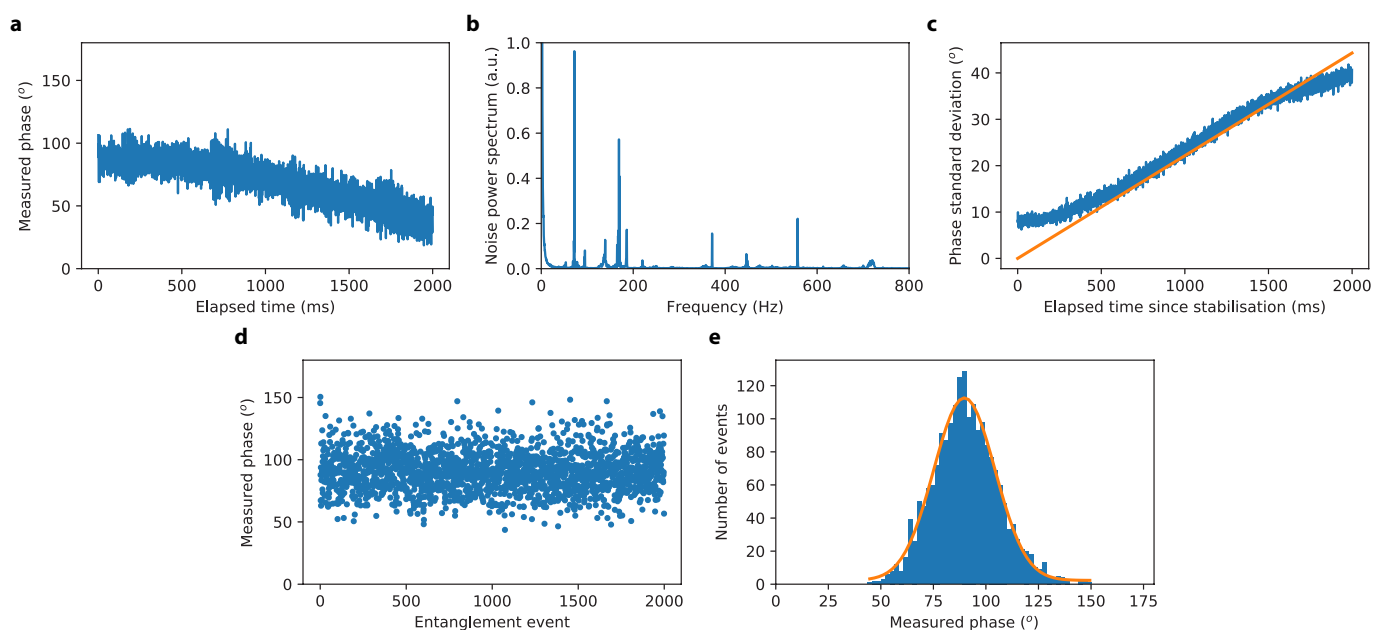
Dashed lines show the model given in the main text (without the infidelity parameter). **b**, This infidelity also applies to the model shown in Fig. 3, because an equally large number of entanglement repetitions was used in generating the data. Error bars for data and shaded model uncertainties are 1 s.d.



**Extended Data Fig. 5 | Verifying passive charge-state stabilization into  $NV^-$ .** **a**, Elementary sequence to probe the NV-centre ionization rate. **b**, Applying our sequence many times results in decay of the NV-centre readout fidelity due to ionization (error bars represent 1 s.d.; lines show fitted exponential decays as guides to the eye). By exploring the ionization

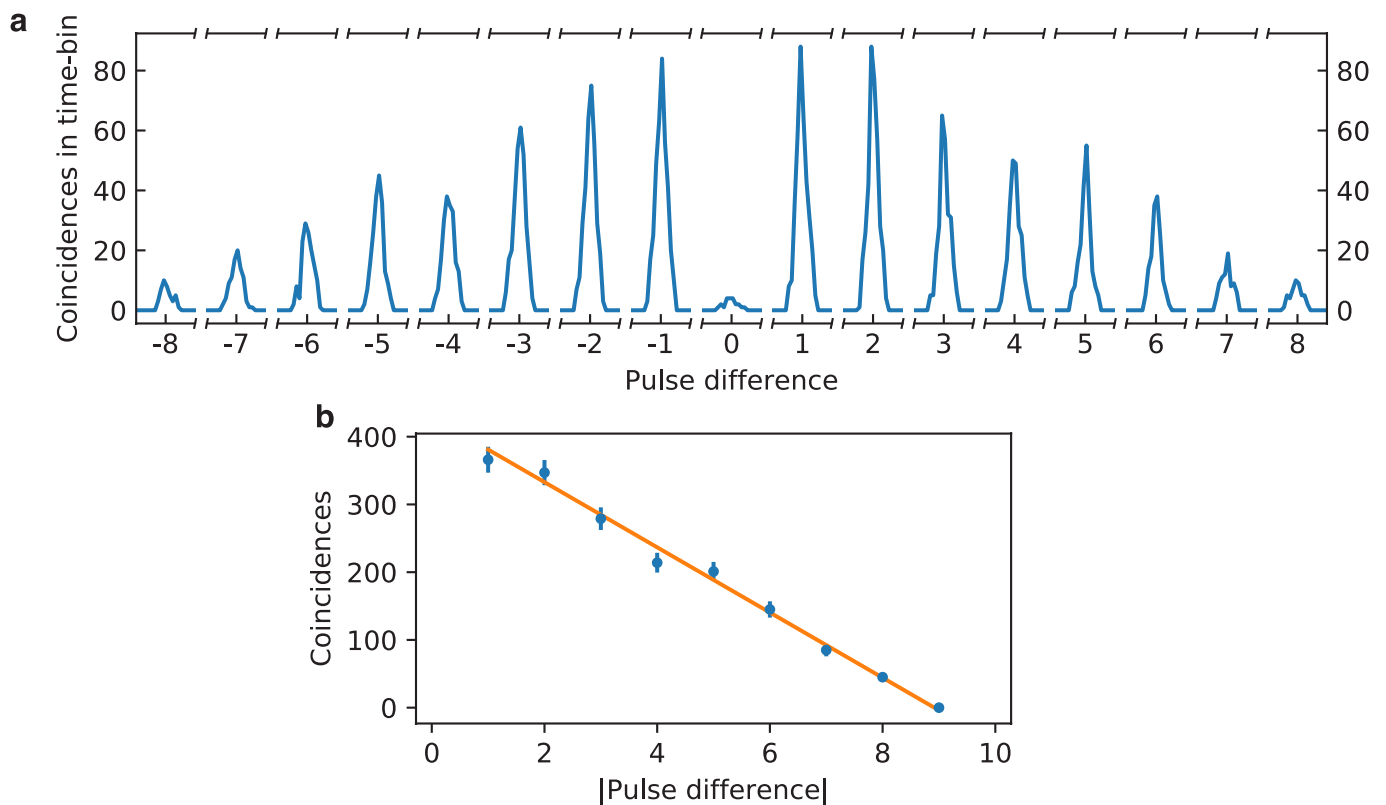
rate for different charge-reset powers, we find an optimal regime in which the spin initialization of  $NV^-$  is not affected by the additional blue-detuned beam and ionization is effectively mitigated over thousands of trials.





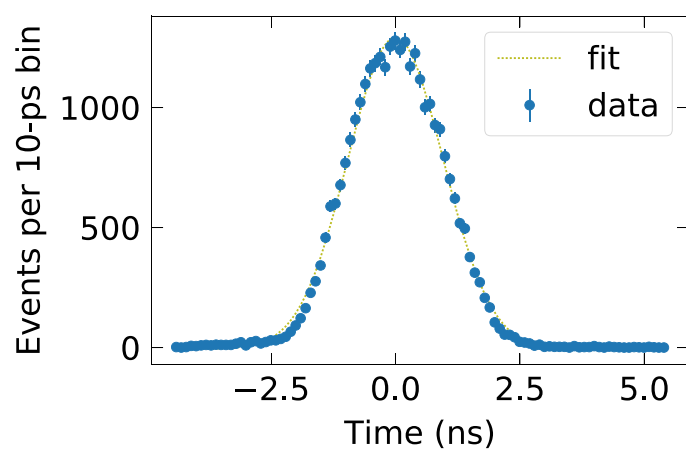
**Extended Data Fig. 6 | Optical-phase stabilization.** Single-photon entanglement requires that the optical phase of an effective interferometer between the two nodes is known. **a**, A typical trace of the interferometer optical phase as it is tracked passively for 2 s. **b**, Power spectrum of the optical-phase signal, showing peaks thought to be due to mechanical resonances of components in the set-up. **c**, Active phase stabilization is used to correct for phase drifts. Here the phase is stabilized and then the interferometer is allowed to passively drift for 2 s. The standard deviation of the phase as a function of time is plotted for a dataset of 100 of these

measurements. The orange line shows a linear fit, used to estimate the rate of phase drift  $\nu_{\text{int}} \approx 20^\circ \text{ s}^{-1}$ . **d**, Here the phase is repeatedly actively stabilized every 180 ms. Entanglement generation occurs during the periods in between stabilization. The interferometer phase is measured directly after each successful heralded entanglement event. **e**, Histogram of the measured post-entanglement optical phases (blue). A Gaussian fit with a standard deviation fixed to the average measured standard deviation for all entanglement data taken,  $\sigma_{\text{int}} = 14.3(3)^\circ$ , is also plotted (orange).



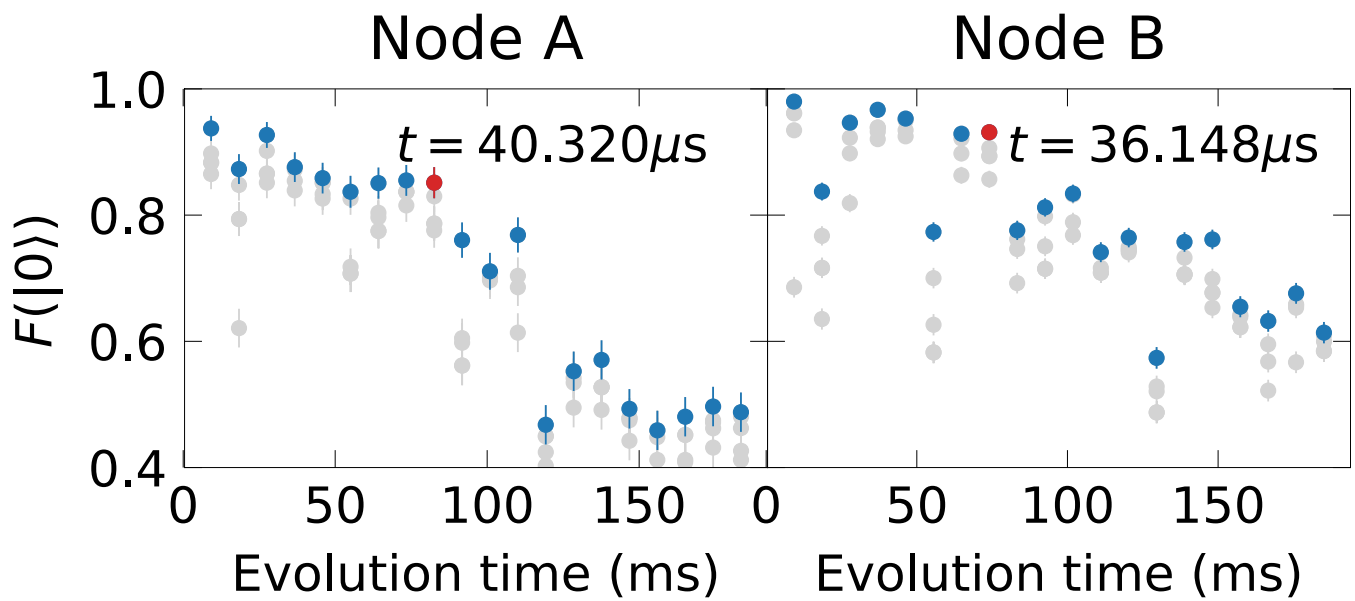
**Extended Data Fig. 7 | Two-photon quantum interference.** **a**, Histogram for coincident events measured by two single-photon detectors in a two-photon quantum interference experiment, measured by cross-referencing photon detection events from a pulse train of 10 optical  $\pi$  pulses that excite both emitters. Hong–Ou–Mandel interference of simultaneously coinciding photons ideally results in vanishing coincidence events within a single excitation round. The time difference between individual excitation rounds is 1  $\mu$ s. Histograms of coincidence counts are shown with a bin-size

of 5 ns. **b**, Total number of coincidences as a function of the number of pulses separating the two detection events. We extrapolate the measured coincidences to infer the expected coincidences for distinguishable photons at zero pulse difference by fitting a linear regression (orange). Using this to normalize the 22 observed coincidences for zero pulse difference allows us to estimate the two-photon quantum interference visibility  $V = 0.90(2)$ . Error bars are 1 s.d.



**Extended Data Fig. 8 | Width of the optical  $\pi$  pulse.** We find a full-width at half-maximum of 2.26 ns. This measurement is necessary to compute the dual-excitation probability (Extended Data Table 1), given a radiative lifetime of 12 ns. Error bars are 1 s.d.





**Extended Data Fig. 9 | Determining the optimal inter-pulse delay for state storage and 1,024 inversion pulses.** We initialize a superposition state on the NV-centre electron spin, preserve it via dynamical decoupling and finally perform optical readout after another  $\pi/2$  pulse. We probe the coherence of the NV centre by varying the inter-pulse delay  $t$  in steps of the Larmor period  $1/\nu_L \approx 2.25 \mu\text{s}$  and shifting the delay in steps of 4 ns for

a total of five data points per Larmor period (grey). For each multiple of the Larmor period we pick the best (the most preserving) inter-pulse delay (blue). We determine the optimal delay  $t$  by selecting an inter-pulse delay that provides sufficient state preservation (about 100 ms) for a moderate number of pulses (red data point and inset text). Left, node A; right, node B. Error bars are 1 s.d.

**Extended Data Table 1 | Independently measured experimental parameters for the performance of the nodes used in our experiment**

	Node A	Node B	Description
$T_2$ (ms)	290(20)	680(70)	Dephasing time of the electron spin state.
$T_1$ (s)	$> 1$	$> 1$	Relaxation time of electron spin eigenstates.
$p_{\text{det}}$ ( $10^{-4}$ )	2.8(1)	4.2(1)	Probability to detect a ZPL photon after a single excitation.
$p_{\text{ionize}}$	$\leq 10^{-6}$	$\leq 10^{-6}$	Probability of passive charge-state control failure per entangling attempt (detailed in methods).
$t$ ( $\mu\text{s}$ )	40.320	36.148	Optimized inter-pulse delay for state storage.
$F_0$	0.959(3)	0.950(3)	Fidelity of the electron read-out for $ \uparrow\rangle$ .
$F_{\pm 1}$	0.995(1)	0.996(1)	Fidelity of the electron read-out for $ \downarrow\rangle$ .
$V$	0.90(2)		Visibility of the two-photon quantum interference (detailed in methods).
$p_{2\text{ph}}$	0.04		Estimated probability of double excitation during the optical $\pi$ -pulse (detailed in methods).
$\nu_{\text{dark}}$ (Hz)	20		Dark count rate per detection channel.
$\sigma_{\text{Int}}$	14.3(1) $^\circ$		Initial uncertainty of the interferometric drift (detailed in methods).
$\nu_{\text{Int}}$ (/s)	$\sim 20^\circ$		Estimated drift rate of the free running interferometer (detailed in methods).

# Printing ferromagnetic domains for untethered fast-transforming soft materials

Yoonho Kim<sup>1,2,5</sup>, Hyunwoo Yuk<sup>1,5</sup>, Ruike Zhao<sup>1,5</sup>, Shawn A. Chester<sup>3</sup> & Xuanhe Zhao<sup>1,4\*</sup>

Soft materials capable of transforming between three-dimensional (3D) shapes in response to stimuli such as light, heat, solvent, electric and magnetic fields have applications in diverse areas such as flexible electronics<sup>1,2</sup>, soft robotics<sup>3,4</sup> and biomedicine<sup>5–7</sup>. In particular, magnetic fields offer a safe and effective manipulation method for biomedical applications, which typically require remote actuation in enclosed and confined spaces<sup>8–10</sup>. With advances in magnetic field control<sup>11</sup>, magnetically responsive soft materials have also evolved from embedding discrete magnets<sup>12</sup> or incorporating magnetic particles<sup>13</sup> into soft compounds to generating nonuniform magnetization profiles in polymeric sheets<sup>14,15</sup>. Here we report 3D printing of programmed ferromagnetic domains in soft materials that enable fast transformations between complex 3D shapes via magnetic actuation. Our approach is based on direct ink writing<sup>16</sup> of an elastomer composite containing ferromagnetic microparticles. By applying a magnetic field to the dispensing nozzle while printing<sup>17</sup>, we reorient particles along the applied field to impart patterned magnetic polarity to printed filaments. This method allows us to program ferromagnetic domains in complex 3D-printed soft materials, enabling a set of previously inaccessible modes of transformation, such as remotely controlled auxetic behaviours of mechanical metamaterials with negative Poisson's ratios. The actuation speed and power density of our printed soft materials with programmed ferromagnetic domains are orders of magnitude greater than existing 3D-printed active materials. We further demonstrate diverse functions derived from complex shape changes, including reconfigurable soft electronics, a mechanical metamaterial that can jump and a soft robot that crawls, rolls, catches fast-moving objects and transports a pharmaceutical dose.

Our composite ink for 3D printing consists of magnetizable microparticles of neodymium–iron–boron (NdFeB) alloy (Extended Data Fig. 1a) and fumed silica nanoparticles (Extended Data Fig. 1b) embedded in a silicone rubber matrix containing silicone catalyst and crosslinker (Fig. 1a). The fumed silica within the silicone resin serves as a rheological modifier to induce the mechanical properties required for direct ink writing<sup>3,16</sup> including shear thinning (Extended Data Fig. 2a) and shear yielding (Extended Data Fig. 2b). These properties ensure that the composite ink can be extruded through a micro-nozzle when pressurized and that the deposited inks maintain their shapes even when stacked up to form multiple layers. The composite ink is prepared first by mixing the non-magnetized NdFeB particles and the silica nanoparticles with the uncured elastomer matrix and then magnetized to saturation under an impulse field (about 2.7 T). The presence of yield stress in the composite ink helps to prevent the dispersed magnetized particles from agglomerating (Extended Data Fig. 3a).

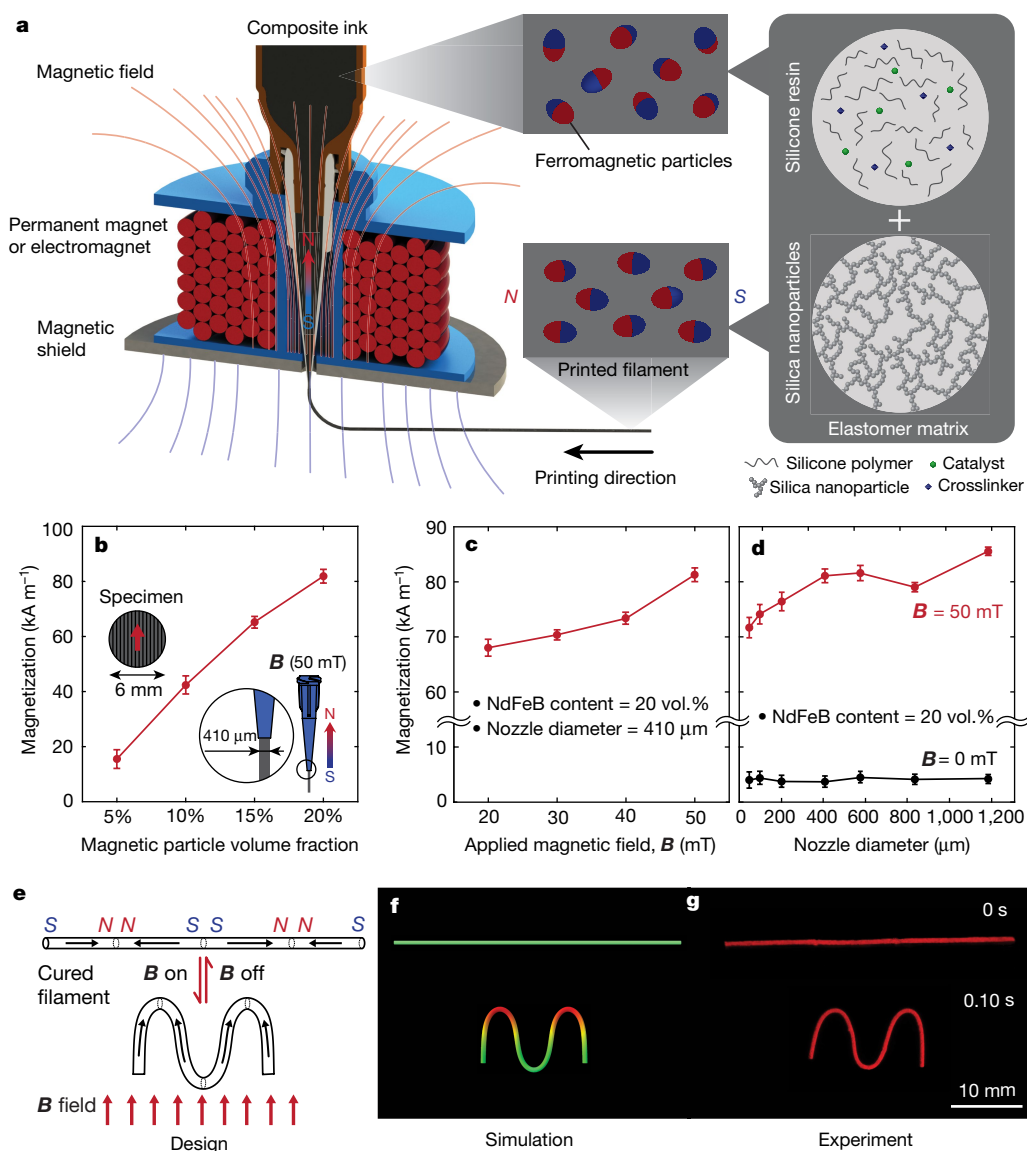
During the printing process, a magnetic field is applied along (or in reverse to) the flow direction of the ink via a permanent magnet or an electromagnetic coil placed around the dispensing nozzle (Fig. 1a). The applied field makes the magnetized NdFeB particles reorient along the field direction, imparting a permanent magnetic moment to the extruded ink filament. The magnetic polarities of the deposited inks

can be tuned either by switching the applied field direction or changing the printing direction. Using this approach, a 3D structure can be encoded with intricate patterns of ferromagnetic domains depending on the magnetic polarities of the filaments that are arranged to construct the 3D structure. To avoid interference in the programmed domains of the printed structure by the applied field at the nozzle, a magnetic shield is used to attenuate the magnetic flux density under the nozzle tip (Fig. 1a). When the printing process is complete, the printed structure is cured at 120 °C for 1 h, during which the presence of yield stress in the uncured ink helps the programmed ferromagnetic domains to remain unaffected by thermal randomization of the aligned particles.

To evaluate the efficacy of our method in printing ferromagnetic domains, we measure the magnetic moment density, or magnetization, in samples (Extended Data Fig. 4) printed under various conditions including the magnetic particle content, the applied field strength and the nozzle diameter. First, samples are printed with magnetic inks containing different volume fractions of NdFeB particles through a nozzle with diameter 410  $\mu\text{m}$  under a magnetic field of 50 mT at the nozzle tip. The measured magnetic moment density varies almost linearly from 16  $\text{kA m}^{-1}$  to 81  $\text{kA m}^{-1}$  as the volume fraction of NdFeB particles in the composite ink increases from 5% to 20% (Fig. 1b). Next, as the applied field at the nozzle tip increases from 20 mT to 50 mT, the magnetic moment density of printed samples (with 20 vol% NdFeB particles through nozzles with diameter 410  $\mu\text{m}$ ) increases from 68  $\text{kA m}^{-1}$  to 81  $\text{kA m}^{-1}$  (Fig. 1c). When the nozzle diameter varies from 200  $\mu\text{m}$  to 1,190  $\mu\text{m}$ , the magnetic moment density of printed samples (with 20 vol% NdFeB particles under a magnetic field of 50 mT at the nozzle tip) increases from 76.6  $\text{kA m}^{-1}$  to 85.4  $\text{kA m}^{-1}$  (Fig. 1d). When printed with very fine nozzles (with diameters of 50  $\mu\text{m}$  and 100  $\mu\text{m}$ ), the fibre diameter becomes larger than the nozzle diameter owing to the die-swelling effect (Extended Data Fig. 1c, d). The ratio between the fibre and nozzle diameters decreases as the nozzle diameter increases, reaching almost one when the nozzle diameter is larger than 200  $\mu\text{m}$  (Extended Data Fig. 1e–g). Printing in the absence of external magnetic fields yields magnetization values below 5  $\text{kA m}^{-1}$  for all nozzle diameters, because the particles are randomly oriented. Furthermore, we print samples in the absence of external field and then magnetize them under impulse fields (about 2.7 T) after curing, which yields the maximum achievable magnetic moment density at each volume fraction of NdFeB particles. In comparison, printing under magnetic fields of 50 mT yields a magnetic moment density that corresponds to about 63–64% of the maximum achievable value at the same concentration of NdFeB particles (Extended Data Fig. 5).

We develop a model to predict the transformation of complex 3D-printed structures with programmed ferromagnetic domains under magnetic fields. Application of magnetic fields induces torques on the embedded ferromagnetic particles, which create stresses that collectively lead to a macroscale material response. If the magnetic moment density (magnetization), a vector quantity, is  $\mathbf{M}$  at a certain point of an incompressible body in the reference configuration, the magnetization vector at the same material point in the deformed body

<sup>1</sup>Soft Active Materials Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>2</sup>Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>3</sup>Department of Mechanical and Industrial Engineering, New Jersey Institute of Technology, Newark, NJ, USA. <sup>4</sup>Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>5</sup>These authors contributed equally: Yoonho Kim, Hyunwoo Yuk, Ruike Zhao. \*e-mail: xzhao@mit.edu



**Fig. 1 | Design of ferromagnetic domains in 3D-printed soft materials.** **a**, Schematics of the printing process and the material composition. The ferromagnetic particles embedded in the composite ink are reoriented by the applied magnetic field generated by a permanent magnet or an electromagnet placed around the dispensing nozzle. **b**, Effect of the volume fraction of magnetized NdFeB particles in the ink on the magnetization of printed samples. **c**, Effect of the applied field strength around the nozzle on the magnetization of printed samples. **d**, Effect of nozzle diameter on the magnetization of printed samples. Samples printed in the absence of applied magnetic fields give magnetization values below  $5 \text{ kA m}^{-1}$ . Error

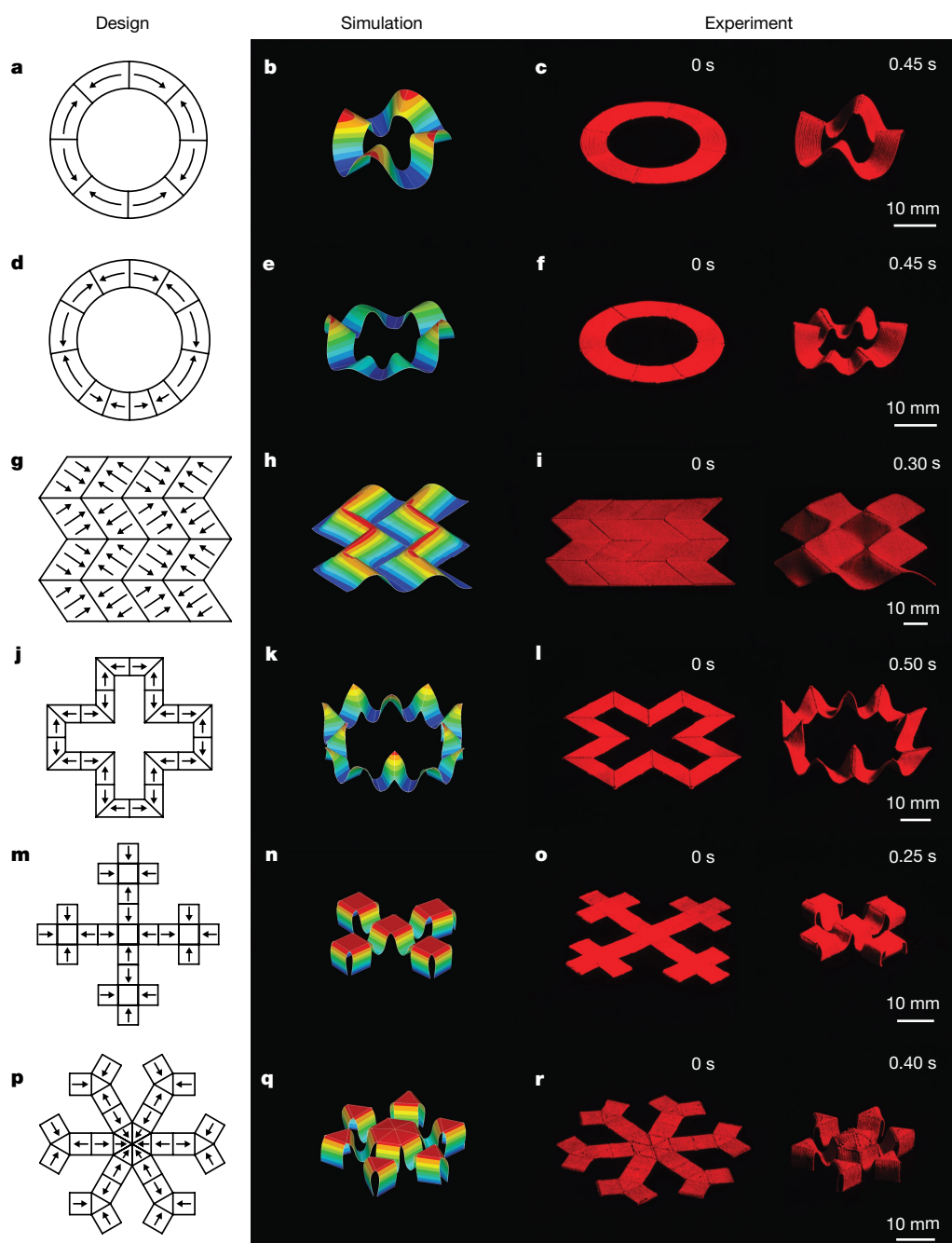
bars indicate the standard deviation for  $n = 3$  measurements at each data point. **e–g**, Schematic illustration (**e**), simulation of the finite-element model (**f**) and experimental results (**g**) of a single fibre encoded with alternating magnetic domains designed to form an 'm' shape in 0.1 s under an applied magnetic field of 200 mT. The elastomeric fibre is printed using a nozzle of diameter 840  $\mu\text{m}$  while switching the direction of magnetic fields (50 mT at the nozzle tip) generated by an electromagnetic coil that encompasses the nozzle. All samples discussed in **c**, **d** and **g** are prepared with the elastomeric composite ink containing 20 vol.% of magnetized NdFeB particles.

can be expressed as  $\mathbf{FM}$ , where  $\mathbf{F}$  denotes the deformation gradient tensor at the point. Then, the magnetic potential energy per unit reference volume under an applied magnetic field  $\mathbf{B}$  can be expressed as  $W^{\text{magnetic}} = -\mathbf{FM} \cdot \mathbf{B}$ , under the assumptions that the presence of soft materials does not substantially alter the applied field and that the potential energy from higher-order terms of  $\mathbf{B}$  and  $\mathbf{M}$  are negligible. From the magnetic potential energy density, the Cauchy stress tensor induced by the applied field on the magnetic moments can be calculated as  $\sigma^{\text{magnetic}} = -\mathbf{B} \otimes \mathbf{FM}$ , where the operation  $\otimes$  denotes the dyadic product, which takes two vectors to yield a second-order tensor. To simulate the deformation of complex structures programmed with ferromagnetic domains, the magnetic stress tensor is implemented as a user-defined element subroutine in the commercial finite-element analysis software ABAQUS (details are available in Supplementary Information).

As an illustrative example to demonstrate the ability to program ferromagnetic domains, a straight filament is printed with an alternating magnetization pattern as illustrated in Fig. 1e by switching the applied field direction during the printing. Upon application of a uniform magnetic field of 200 mT, the straight filament transforms into an 'm' shape in 0.1 s (Fig. 1g), and quickly reverts to its original shape upon removal of the applied field in 0.2 s. Such rapid, reversible transformation can be repeated on demand by magnetic actuation (Supplementary Video 1). The simulation conducted under the same conditions, including the magnetic and mechanical properties and the applied field as in the experiment, is in good agreement with the experimental results (Fig. 1f), validating the use of model-based simulation to guide the design of complex shape-morphing structures with programmed ferromagnetic domains.

In Fig. 2, we present a set of two-dimensional planar structures that rapidly transform into complex 3D shapes under the applied magnetic





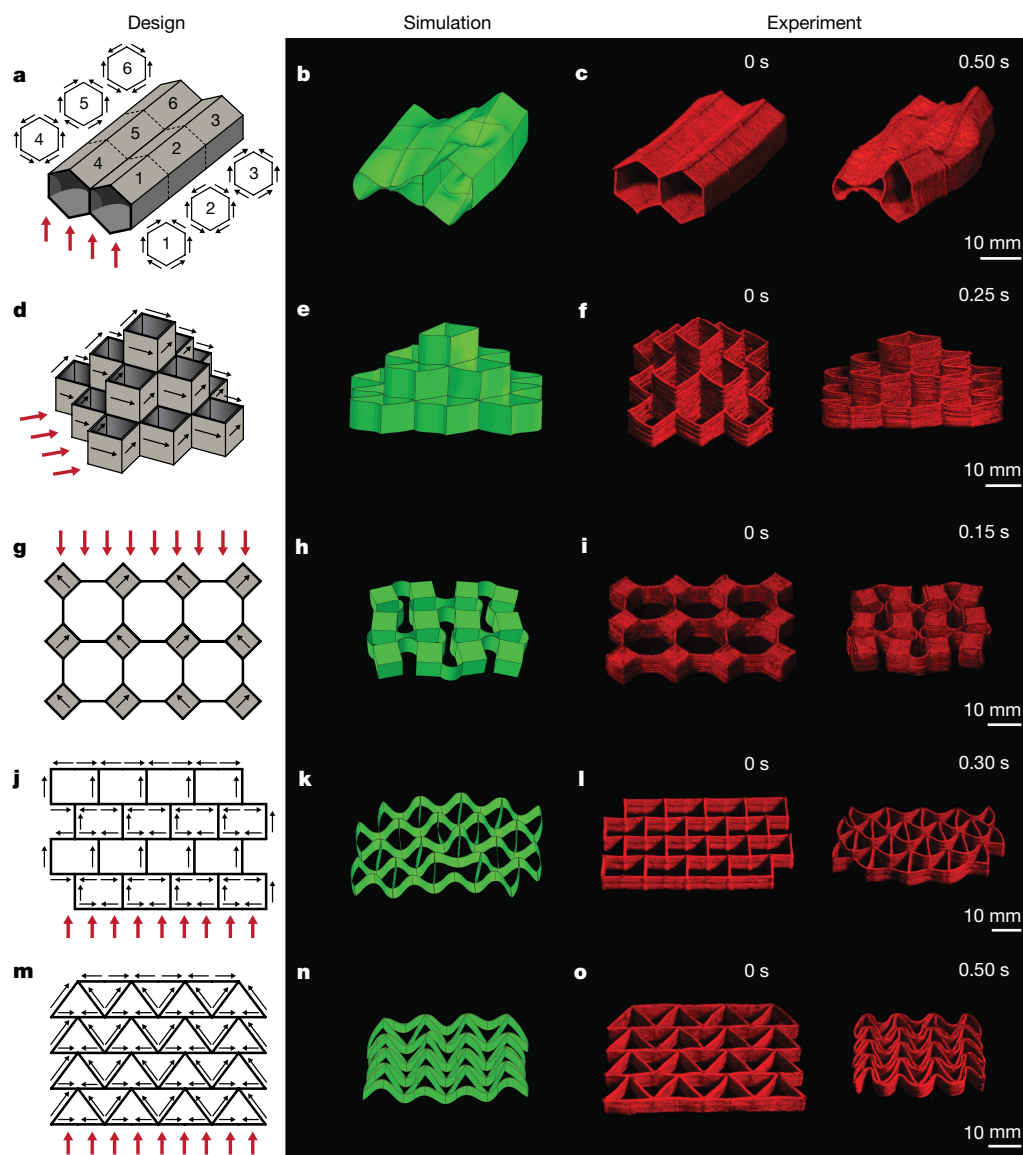
**Fig. 2 | Various two-dimensional planar structures with programmed ferromagnetic domains demonstrating complex shape changes under applied magnetic fields.** **a–r**, Schematic designs, finite-element simulations and experimental results for an annulus encoded with alternating domains that are equidistant (**a–c**); an annulus encoded with alternating domains that vary in size (**d–f**); a Miura-ori fold encoded with alternating oblique patterns of ferromagnetic domains (**g–i**); a hollow cross encoded with alternating ferromagnetic domains along the perimeter (**j–l**); quadrupedal (**m–o**) and hexapedal (**p–r**) structures enabled by

folding of the magnetically active segments surrounding the magnetically inactive segments (unlabelled areas in the schematic designs). All of the demonstrated structures are printed with the elastomeric composite ink containing 20 vol% of magnetized NdFeB particles using a nozzle of diameter 410  $\mu\text{m}$  under a magnetic field of 50 mT at the nozzle tip generated by a permanent magnet. Actuation of the demonstrated structures is performed by applying magnetic fields of 200 mT perpendicular to the planes of the structures. The detailed dimensions of the printed structures are given in Extended Data Fig. 6a–f.

fields of 200 mT as a result of the programmed ferromagnetic domains. In Fig. 2a and d, we design two annular rings with the same geometry but different patterns of ferromagnetic domains to illustrate the effects of programmed domains on the macroscale response. Our model-based simulation predicts that the two rings should yield different 3D morphologies under the same magnetic field applied perpendicularly to their planes. The second annulus encoded with alternating patterns that vary in magnitude gives a more complex undulating shape (Fig. 2e) than does the first annulus (Fig. 2b), whose alternating patterns are

equidistant. The simulation results are in good agreement with experimental results (Fig. 2c, f, Supplementary Video 2), further demonstrating that our model is capable of guiding the design of complex shape-morphing structures using programmed ferromagnetic domains.

When programmed with more intricate domain patterns, even a simple geometry can yield a complex 3D shape under an applied magnetic field. As an example, in Fig. 2g, we design a simple rectangular structure with alternating oblique patterns of ferromagnetic domains to create a Miura-ori pattern<sup>18</sup>. This untethered structure provides fast



**Fig. 3 | Various 3D structures with programmed ferromagnetic domains demonstrating complex shape changes under applied magnetic fields. a–o**, Schematic designs, finite-element simulations and experimental results for two adjoining hexagonal tubes programmed to form undulating surfaces under the applied magnetic field owing to the alternating ferromagnetic domains (a–c); a pyramid-shaped thin-walled structure exhibiting elongation in its diagonal direction along the applied magnetic field (d–f); a set of auxetic structures (with negative Poisson's

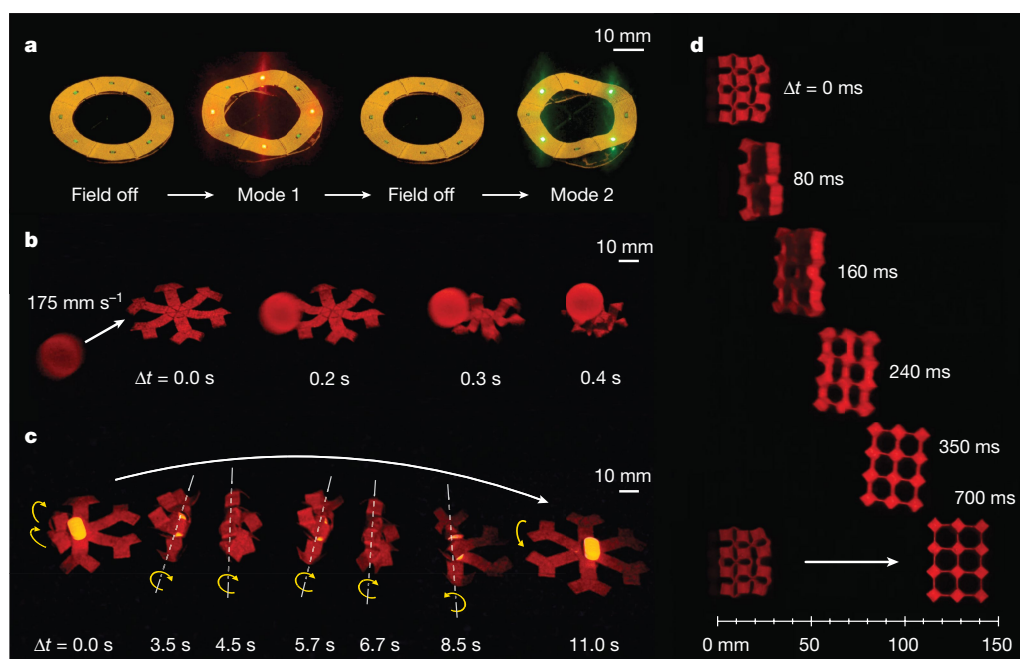
ratios) exhibiting shrinkage in both length and width under applied magnetic fields (g–o). All of the demonstrated structures are printed with elastomeric composite ink containing 20 vol% of magnetized NdFeB particles using a nozzle of diameter 410  $\mu\text{m}$  under a magnetic field of 50 mT at the nozzle tip, generated by a permanent magnet. Actuation of the demonstrated structures is performed by applying magnetic fields of 200 mT along the directions indicated in a, d, g, j and m. The detailed dimensions of the printed structures are given in Extended Data Fig. 6g–k.

(in 0.3 s) and fully reversible folding and unfolding under magnetic actuation (Fig. 2i and Supplementary Video 2), as predicted by our model (Fig. 2h). Notably, the response speed of our Miura-ori structure with programmed ferromagnetic domains is much faster than existing ones in the literature based on liquid crystal elastomers<sup>19,20</sup>, shape memory polymers<sup>21,22</sup> and thermally responsive hydrogels<sup>23</sup>.

When more intricate designs are programmed with ferromagnetic domains, as illustrated in Fig. 2j, m and p, the two-dimensional planar structures transform into more complex 3D shapes (Fig. 2l, o, r and Supplementary Video 2); for these it is no longer straightforward to trace the original shapes without knowing the programmed domains. Design and realization of such complex shape-morphing structures are enabled by the model-based simulations, which accurately predict the complex 3D morphologies (Fig. 2k, n, q and Supplementary Video 2). Previously, a transition from two-dimensional planar structures to complex 3D shapes has been achieved by controlled buckling of materials that are selectively attached on biaxially stretched

elastomeric substrates<sup>24,25</sup>. Compared with existing techniques, our method based on printing ferromagnetic domains offers additional advantages in two-dimensional to 3D structural transition, including (i) substrate-free, remote actuation, (ii) fast and fully reversible transformation and (iii) the capability to selectively actuate specific parts of the structure.

Our method of printing ferromagnetic domains can be further extended to complex 3D structures. When printing 3D structures with direct ink writing, however, difficulties typically arise owing to structural instability as the deposited filaments are stacked up. To ensure a more stable printing process, we introduce a support ink composed of a silicone resin containing catalyst and fumed silica nanoparticles (Extended Data Fig. 7; see Extended Data Fig. 2a and b for rheological properties). When printed, the support ink serves as a fugitive support that buttresses the adjacent magnetic ink (Extended Data Fig. 7a and Supplementary Video 3). After the magnetic ink is fully cured, the support ink can be removed by solvent rinses (Extended Data Fig. 7c).



**Fig. 4 | Functional demonstrations of 3D-printed soft materials with programmed ferromagnetic domains.** **a**, A reconfigurable soft electronic device (as detailed in Extended Data Fig. 9) based on the annular ring structure exhibiting different electronic functions depending on the direction of an applied magnetic field of 30 mT. **b**, A hexapedal structure stopping and holding a fast-moving object (glass ball of diameter 18 mm and weight 8 g) upon application of a magnetic field generated

by a permanent magnet. **c**, A hexapedal structure wrapping an oblong pharmaceutical pill and carrying the pill using rolling-based locomotion under a rotating magnetic field generated by a permanent magnet. **d**, Horizontal leap of a 3D auxetic structure upon sudden reversal of the applied magnetic field direction while attenuating the field strength by rotating a permanent magnet by 90°. Detailed information on how to apply the magnetic fields to achieve actuation is given in Extended Data Fig. 10.

The use of support ink and the consequent ability to print 3D structures with programmed domains allow us to create a set of high-aspect-ratio multilayered structures (Fig. 3) that exhibit rapid and reversible transformation between complex 3D shapes under magnetic fields of 200 mT. In Fig. 3c, we present a thin-walled structure consisting of two adjoining hexagonal tubes with high aspect ratios. The ferromagnetic domains are programmed in such a way that some parts of the tubes expand while the others collapse, as illustrated in Fig. 3a, to create complex undulating surfaces in a continuous 3D structure under the applied magnetic field, as predicted and observed by the simulation and the experiment (Fig. 3b, c and Supplementary Video 4), respectively. In another example, to demonstrate the versatility of our fabrication method, we create a pyramid-shaped thin-walled structure that elongates along the direction of applied magnetic fields (Fig. 3e, f and Supplementary Video 4) as a result of the programmed magnetic domains (Fig. 3d).

The versatility of our model-guided design and fabrication method enables us to create auxetic structures (Fig. 3i, l, o), a type of mechanical metamaterials characterized by negative Poisson's ratios. Our printed auxetic structures exhibit shrinkage in both length and width in response to external magnetic fields. Typically, mechanical metamaterials show auxetic behaviours only when uniaxially compressed or stretched and thus require direct mechanical contact<sup>26</sup>. In addition, owing to the limited fabrication techniques available to achieve complex designs, remote actuation of untethered auxetic structures has not been realized in other types of active materials. Guided by our model-based predictions (Fig. 3h, k, n), we design a set of mechanical metamaterials with programmed ferromagnetic domains (Fig. 3g, j, m) that quickly shrink in both length and width under the applied fields within 0.5 s and recover their original shapes upon removal of the applied fields (Fig. 3i, l, o and Supplementary Video 5). The use of magnetic fields as an actuation method obviates the need for direct contact in realizing auxetic behaviours in mechanical metamaterials.

In the design and fabrication of shape-programmable soft materials, intensive efforts have been made to increase the level of complexity by

adopting 3D printing techniques such as inkjet printing<sup>27</sup>, stereolithography<sup>28,29</sup> and direct ink writing<sup>30</sup>. However, fast and fully reversible actuation between programmed shapes has remained a central challenge in the field. To quantitatively evaluate the actuation performance, we compare the energy density and the actuation rate (Extended Data Fig. 8a) of printed shape-programmable materials in the literature. We also compare the power density (Extended Data Fig. 8b), one often-used metric to evaluate the actuation performance of active materials. Our shape-morphing structures shown in Figs. 2 and 3 deform up to strain levels from 0.15 to 0.25 within 0.1 s to 0.5 s, providing a power density ranging from 22.3 kW m<sup>-3</sup> to 309.3 kW m<sup>-3</sup>, which are orders of magnitude greater than the actuation rates and power densities achieved by existing 3D-printed shape-transforming soft materials.

The capability to create complex shape changes allows us to achieve diverse functions from our printed structures, as shown in Fig. 4. First, by combining electronic components and circuitry with our annular ring structure in Fig. 2c, we print a soft electronic device as detailed in Extended Data Fig. 9a. This soft electronic device deforms into two different shapes depending on the direction of applied magnetic fields of 30 mT, and each mode of transformation yields a different electronic function (Fig. 4a, Extended Data Fig. 9b, c and Supplementary Video 6). The results demonstrate that our multimaterial 3D printing method gives functionally reconfigurable soft electronic devices, whose rigid-material counterparts have recently been achieved by means of multistable buckling<sup>31</sup>.

We further demonstrate the capability of interacting with an object based on the complex shape changes of the hexapedal structure shown in Fig. 2r. Using the fast response upon magnetic actuation, the hexapedal structure quickly stops a fast-moving object (Fig. 4b and Supplementary Video 7). When applying a magnetic field in the opposite direction to create a reversed shape, the hexapedal structure can catch a falling object and hold it against external disturbance and then release the object on demand by using the previous mode of transformation (Supplementary Video 7). When a rotating magnetic field is applied, the hexapedal structure can roll up its body and move



forwards and backwards by rolling-based locomotion (Supplementary Video 8). Harnessing the shape changes and motion, the hexapedal structure can carry an object with arbitrary shape such as a round or oblong pharmaceutical pill (Fig. 4c) and release the pill on demand (Supplementary Video 8).

The 3D mechanical metamaterial presented in Fig. 3i can show a horizontal leap based on the drastic release of the elastic and magnetic potential energy (Fig. 4d and Supplementary Video 9). The fast response of the auxetic structure generates an average speed of  $250 \text{ mm s}^{-1}$  during the leap, allowing it to move forwards by 120 mm within 0.7 s on the horizontal plane. This leaping motion is achieved by first applying a magnetic field in one direction to collapse the auxetic structure and then switching to a field in the opposite direction while attenuating the field strength. This sudden reversal of the field direction quickly increases the magnetic potential energy and triggers the drastic release of the stored elastic and magnetic potential energy, which is converted to kinetic energy during the horizontal leap.

Our printing method as a fabrication platform can be extended to multiple composite inks using different types of elastomer and hydrogel matrices and magnetic particles. By printing ferromagnetic domains in soft materials, we introduce new design parameters—domain patterns, magnetization strength and actuation fields—into the design and fabrication of shape-programmable soft materials. The remote actuation of such untethered, complex and fast shape-shifting soft materials based on magnetic fields suggests new possibilities for applications in flexible electronics, biomedical devices and soft robotics.

### Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0185-0>.

Received: 22 November 2017; Accepted: 11 April 2018;

Published online 13 June 2018.

- Ma, M., Guo, L., Anderson, D. G. & Langer, R. Bio-inspired polymer composite actuator and generator driven by water Gradients. *Science* **339**, 186–189 (2013).
- Zarek, M. et al. 3D printing of shape memory polymers for flexible electronic devices. *Adv. Mater.* **28**, 4449–4454 (2016).
- Wehner, M. et al. An integrated design and fabrication strategy for entirely soft, autonomous robots. *Nature* **536**, 451–455 (2016).
- Park, S. J. et al. Phototactic guidance of a tissue-engineered soft-robotic ray. *Science* **353**, 158–162 (2016).
- Zhao, X. H. et al. Active scaffolds for on-demand drug and cell delivery. *Proc. Natl Acad. Sci. USA* **108**, 67–72 (2011).
- Fusco, S. et al. An integrated microrobotic platform for on-demand, targeted therapeutic interventions. *Adv. Mater.* **26**, 952–957 (2014).
- Davis, K. A., Burke, K. A., Mather, P. T. & Henderson, J. H. Dynamic cell behavior on shape memory polymer substrates. *Biomaterials* **32**, 2285–2293 (2011).
- Erb, R. M., Martin, J. J., Soheilani, R., Pan, C. & Barber, J. R. Actuating soft matter with magnetic torque. *Adv. Funct. Mater.* **26**, 3859–3880 (2016).
- Hines, L., Petersen, K., Lum, G. Z. & Sitti, M. Soft actuators for small-scale robotics. *Adv. Mater.* **29**, 1603483 (2017).
- Martel, S. Beyond imaging: macro- and microscale medical robots actuated by clinical MRI scanners. *Science Robotics* **2**, eaam8119 (2017).
- Rahmer, J., Stehning, C. & Gleich, B. Spatially selective remote magnetic actuation of identical helical micromachines. *Science Robotics* **2**, eaal2845 (2017).
- Boncheva, M. et al. Magnetic self-assembly of three-dimensional surfaces from planar sheets. *Proc. Natl Acad. Sci. USA* **102**, 3924–3929 (2005).
- Kim, J., Chung, S., Choi, S., Lee, H. & Kwon, S. Programming magnetic anisotropy in polymeric microactuators. *Nat. Mater.* **10**, 747–752 (2011).
- Lum, G. Z. et al. Shape-programmable magnetic soft matter. *Proc. Natl Acad. Sci. USA* **113**, E6007–E6015 (2016).
- Hu, W., Lum, G. Z., Mastrangeli, M. & Sitti, M. Small-scale soft-bodied robot with multimodal locomotion. *Nature* **554**, 81–85 (2018).
- Lewis, J. A. Direct ink writing of 3D functional materials. *Adv. Funct. Mater.* **16**, 2193–2204 (2006).
- Kokkinis, D., Schaffner, M. & Studart, A. R. Multimaterial magnetically assisted 3D printing of composite materials. *Nat. Commun.* **6**, 8643 (2015).
- Silverberg, J. L. et al. Using origami design principles to fold reprogrammable mechanical metamaterials. *Science* **345**, 647–650 (2014).
- Yuan, C. et al. 3D printed reversible shape changing soft actuators assisted by liquid crystal elastomers. *Soft Matter* **13**, 5558–5568 (2017).
- Ware, T. H., McConney, M. E., Wie, J. J., Tondiglia, V. P. & White, T. J. Voxelated liquid crystal elastomers. *Science* **347**, 982–984 (2015).
- Oyefusi, A. & Chen, J. Reprogrammable chemical 3D shaping for origami, kirigami, and reconfigurable molding. *Angew. Chem.* **129**, 8362–8365 (2017).
- Zhao, Z. et al. Origami by frontal photopolymerization. *Sci. Adv.* **3**, e1602326 (2017).
- Na, J. H. et al. Programming reversibly self-folding origami with micropatterned photo-crosslinkable polymer trilayers. *Adv. Mater.* **27**, 79–85 (2015).
- Xu, S. et al. Assembly of micro/nanomaterials into complex, three-dimensional architectures by compressive buckling. *Science* **347**, 154–159 (2015).
- Zhang, Y. H. et al. A mechanically driven form of kirigami as a route to 3D mesostructures in micro/nanomembranes. *Proc. Natl Acad. Sci. USA* **112**, 11757–11764 (2015).
- Babaei, S. et al. 3D soft metamaterials with negative Poisson's ratio. *Adv. Mater.* **25**, 5044–5049 (2013).
- Ding, Z. et al. Direct 4D printing via active composite materials. *Sci. Adv.* **3**, e1602890 (2017).
- Kim, J., Hanna, J. A., Byun, M., Santangelo, C. D. & Hayward, R. C. Designing responsive buckled surfaces by halftone gel lithography. *Science* **335**, 1201–1205 (2012).
- Ge, Q. et al. Multimaterial 4D printing with tailorable shape memory polymers. *Sci. Rep.* **6**, 31110 (2016).
- Gladman, A. S., Matsumoto, E. A., Nuzzo, R. G., Mahadevan, L. & Lewis, J. A. Biomimetic 4D printing. *Nat. Mater.* **15**, 413–418 (2016).
- Fu, H. et al. Morphable 3D mesostructures and microelectronic devices by multistable buckling mechanics. *Nat. Mater.* **17**, 268–276 (2018).

**Acknowledgements** We thank D. Bono for help in magnetic characterizations. This work is supported by the National Science Foundation (CMMI-1661627) and the Office of Naval Research (N00014-17-1-2920) and the MIT Institute for Soldier Nanotechnologies. Y.K. acknowledges financial support from Harvard-MIT Division of Health Sciences and Technology. H.Y. acknowledges financial support from a Samsung Scholarship.

**Author contributions** Y.K., H.Y., R.Z. and X.Z. designed the study and interpreted the results. H.Y., Y.K. and X.Z. conceived the idea of printing ferromagnetic domains. H.Y. and X.Z. developed the 3D printing platform. Y.K. and H.Y. developed materials and methods of printing and performed material characterizations. Y.K. designed and fabricated the printed structures and demonstrated their functions. X.Z., R.Z. and Y.K. developed the theory for soft materials with ferromagnetic domains, R.Z. and S.A.C. implemented the numerical models, and R.Z. performed the simulations. Y.K., H.Y. and R.Z. produced the figures and videos. Y.K. and X.Z. wrote the manuscript with input from all authors. X.Z. supervised the study.

**Competing interests** The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0185-0>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0185-0>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to X.Z.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## METHODS

**Ink composition and preparation.** The magnetic ink was prepared first by blending two silicone-based materials—SE 1700 (Dow Corning Corp.) and Ecoflex 00-30 Part B (Smooth-on Inc.)—in a 1:2 volume ratio. Ecoflex 00-30 Part B, a softer elastomer than SE 1700, was used to achieve the preferred mechanical properties of the composite material. Fumed silica nanoparticles (amorphous, 20–30 nm; US Research Nanomaterials Inc.), which corresponds to 12.5 wt% with respect to Ecoflex Part B, were added to achieve required rheological properties for direct ink writing. After mixing the blend in a planetary mixer (AR-100, Thinky) at 2,000 r.p.m. for 2 min, 20 vol% NdFeB microparticles (287.5 wt% with respect to Ecoflex Part B) with an average size of 5  $\mu\text{m}$  (MQFP-B-2007609-089, Magnequench) were added into the elastomer mixture and then mixed thoroughly at 2,000 r.p.m. for 3 min, followed by defoaming at 2,200 r.p.m. for 1 min. The composite ink was then magnetized by impulse magnetic fields (about 2.7 T) generated by an impulse magnetizer (IM-10-30, ASC Scientific) to impart magnetic polarities to the ferromagnetic particles embedded in the elastomer matrix. Both SE 1700 and Ecoflex 00-30 are platinum-catalysed, addition-curing silicones, so 10 wt% SE 1700 catalyst with respect to SE 1700 base was added into the magnetized ink and then mixed at 2,000 r.p.m. for 30 s before printing. The final concentrations of components were as follows: 21.78 wt% Ecoflex 00-30 Part B, 2.72 wt% fumed silica nanoparticles, 11.71 wt% SE 1700 base, 1.17 wt% SE 1700 catalyst and 62.62 wt% NdFeB microparticles. For imaging purposes, about 2 wt% fluorescent colourants (Ignite PMS 805C, Smooth-on Inc.) were added to this final composition.

The support ink, which was used for supporting structures when printing multilayered or 3D structures with the magnetic ink, was prepared by mixing a platinum-based silicone-curing accelerator (Elastosil CAT PT-F, Wacker) with fumed silica nanoparticles (amorphous, 20–30 nm; US Research Nanomaterials Inc.) in a 5.45:1 mass ratio. Fumed silica nanoparticles were added to achieve the rheological properties required for direct ink writing of the support ink. The higher concentration of catalyst in the support ink prevents diffusion of catalyst molecules from the adjacent magnetic inks, and therefore helps prevent imperfect curing of the printed magnetic structures. After the magnetic inks were fully cured upon heating at 120 °C for 1 h, the fugitive support ink was removed by rinsing with isopropyl alcohol using an orbital shaker (Micro Plate Shaker, VWR).

**Printing procedure.** The prepared magnetic and support inks were loaded into syringe barrels and defoamed at 2,200 r.p.m. for 1 min to remove trapped air bubbles. The inks were then mounted to the custom-designed 3D printer based on a Cartesian gantry system (AGS1000, Aerotech). Conical nozzles with inner diameter 410  $\mu\text{m}$  (Smoothflow Tapered Tip, Nordson EFD) were used for both inks in our demonstrations (Figs. 2 and 3). The detailed designs including ferromagnetic domain patterns and dimensions of the printed structures in Figs. 2 and 3 are given in Extended Data Fig. 6. The external magnetic fields applied at the nozzle to reorient the magnetic particles embedded in the ink during printing were generated by either an electromagnet or a permanent magnet. Printing paths were generated by CAD drawings (SolidWorks, Dassault Systèmes) and converted into G-code by a commercial software package (CADfusion, Aerotech) and custom Python scripts to command the  $x$ – $y$ – $z$  motion of the printer head. See Supplementary Video 3 for overall printing and actuation processes.

**Rheological characterization.** Rheological responses (Extended Data Fig. 2a, b) of the magnetic and support inks were characterized using a rotational rheometer (AR-G2, TA Instruments) with a 20-mm-diameter steel plate geometry. For magnetic inks, both magnetized and nonmagnetized samples were tested to evaluate the effects of magnetic interaction between the embedded magnetized particles. Apparent viscosities were measured via steady-state flow experiments with a sweep of shear rates (0.01–100  $\text{s}^{-1}$ ). Shear storage moduli were measured as a function of shear stress via oscillation experiments at a fixed frequency of 1 Hz with a sweep of stress (10–10,000 Pa). The magnetic and support inks were equilibrated at 25 °C for 1 min before testing, and all experiments were performed at 25 °C with a gap height of 0.5 mm.

**Magnetic characterization.** The quality of alignment of the magnetic particles was evaluated by measuring the magnetic moment density (magnetization) of the printed samples with a vibrating sample magnetometer (DMS 1660, ADE Technologies). To prepare specimens, a set of parallel lines was printed in the same direction to construct a rectangular film. Then, the printed film was cut into circles using a 6-mm biopsy punch (Millex Inc.) to fit into the sample holder of the machine. The magnetic moments of the samples were measured against a sweep of external magnetic fields from  $-8,000 \text{ A m}^{-1}$  to  $8,000 \text{ A m}^{-1}$ . Remanent magnetization, which corresponds to the measured magnetic moment when the applied external field is zero, was divided by each specimen's volume to obtain the magnetic moment density of the specimen.

**Mechanical testing.** Two types of rectangular planar sheets (width 12 mm, length 35 mm) were printed with an conical nozzle of diameter 840  $\mu\text{m}$  in the absence of external magnetic fields and under applied magnetic fields generated by a permanent magnet around the nozzle, respectively. After curing, the sheets were cut

into dog-bone-shaped specimens with known dimensions (width 4 mm, gauge length 17 mm) for tensile testing. The cross-sectional area of each specimen was calculated by dividing the sample's original volume by its length, where the volume was calculated based on the sample's mass measured before the cut and the density ( $2.434 \text{ g cm}^{-3}$ ) of the composite ink containing 20 vol% of NdFeB. The specimens were tested on a mechanical testing machine (Z2.5, Zwick/Roell) with a 20 N load cell at a strain rate of  $0.01 \text{ s}^{-1}$ . Nominal stress–stretch curves were plotted for both materials, and shear moduli  $\mu$  were obtained by fitting the experimental curves using a neo-Hookean model (Extended Data Fig. 2c). The shear modulus of the magnetized-ink-based material we obtained was  $\mu = 330 \text{ kPa}$ . The specimen printed in the presence of external fields showed higher shear modulus compared with the specimen printed without external fields ( $\mu = 245 \text{ kPa}$ ). This higher shear modulus may be attributed to the field-induced alignment of ferromagnetic particles along the filaments when magnetic fields were applied during the printing.

**Imaging and videography.** Images and videos of the programmed shape changes of the printed samples were taken under a blue LED light source (460-nm wavelength), while applying external magnetic fields in ambient conditions. The external fields were generated by either permanent magnets (K&J Magnetics) or electromagnets (APW Company) depending on the shapes and sizes of the printed samples (Extended Data Fig. 10). Images and videos were taken with a DSLR camera (D7000, Nikon) with a red-coloured filter (HMC R25A, Hoya) that transmits wavelengths above about 600 nm (Supplementary Video 3). To reduce the effects of friction between the printed samples and the substrate, fine glass powders with an average size of 9–13  $\mu\text{m}$  (glass spheres, Sigma Aldrich) were applied to the substrate as a dry lubricant.

**Finite-element analysis.** For all designs presented in Figs. 1–3, the shapes deformed in the presence of external magnetic fields were simulated using a user-defined element subroutine implemented in the commercial finite-element analysis software ABAQUS. For all simulations, the following input parameters were used: the shear modulus  $\mu = 330 \text{ kPa}$ , the bulk modulus  $K = 1,000\mu$  (the large bulk modulus was chosen to approximate incompressibility), and the uniform external magnetic field  $B = 200 \text{ mT}$ . For the magnetization parameter in the simulation, the experimentally measured value ( $M = 81 \text{ kA m}^{-1}$ ; see Fig. 1b) was used for samples printed with inks containing 20 vol% of NdFeB particles through nozzles with 410  $\mu\text{m}$  diameter under the applied field of 50 mT at the nozzle tip.

**Validation of printing-induced magnetization.** The magnetization of a sample printed with a nozzle of diameter 410  $\mu\text{m}$  in the presence of magnetic fields (50 mT) was measured while varying the angular position of the printed fibres with respect to the horizontal direction (Extended Data Fig. 4a). The maximum magnetization value was measured when the printed fibres are aligned with the positive  $x$  direction, in which an external magnetic field is applied by the vibrating sample magnetometer. The measured magnetization value decreased as the angle increased and reached almost zero when the printed fibres were vertically aligned. When the specimen was rotated by 180°, the sign of the measured magnetization was changed, indicating that the specimen's magnetic polarity was reversed, while the magnitude remained almost unchanged (Extended Data Fig. 4b). It demonstrates that the printed fibre direction can represent the overall magnetization direction (that is, magnetic polarity) of the printed sample.

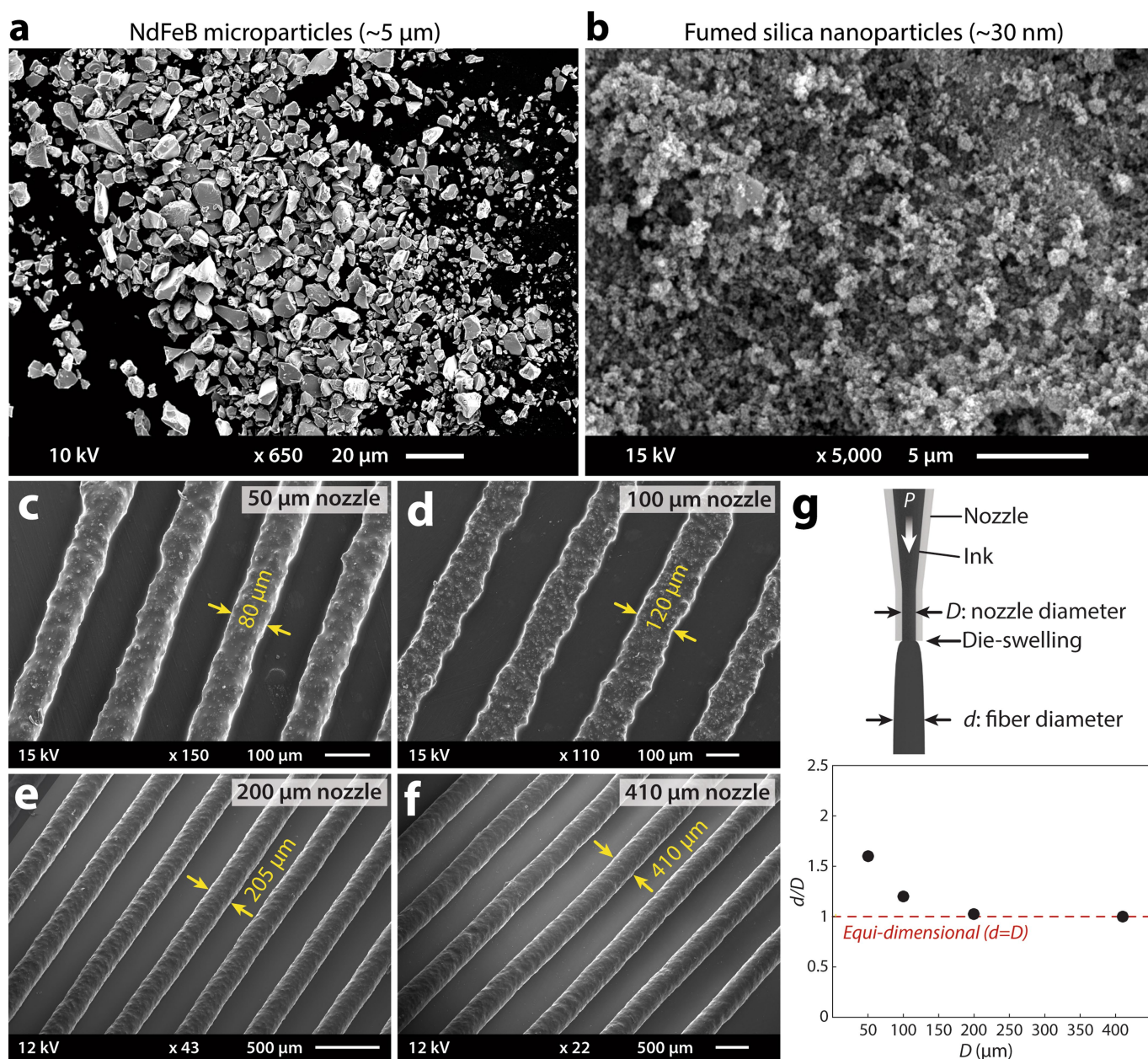
**Calculation of energy and power density.** As evaluation criteria for actuation performance, we used three quantities that can either be found or calculated from the literature. First, we used the actuation rate  $f = 1/t$ , where  $t$  is the time taken to generate one complete actuation cycle. For materials with irreversible actuation such as shape memory polymers<sup>27,32</sup>, the actuation rate was considered to be the time taken for completing the first cycle of actuation. Second, we used the energy density  $\rho_E$ , which can be determined from force–stroke curves for actuation. Third, we used the power density  $\rho_W = f\rho_E$  to evaluate the overall actuation performance of various 3D-printed shape-programmable soft materials. For papers that report force–stroke curves for actuation or the corresponding values of energy density, we used the reported values. When the data were not available, we approximated the values for different classes of materials.

For elastically actuated materials such as shape memory polymers, liquid crystal elastomers and other composite materials, the energy density was approximated as<sup>33</sup>  $\rho_E = E(\varepsilon_a)^2/2$  where  $E$  is Young's modulus in the rubbery state and  $\varepsilon_a$  is the actuation strain. For osmotically actuated materials such as hydrogels and other swelling-based actuators, the energy density was approximated as<sup>34</sup>  $\rho_E = \Delta\Pi\varepsilon_d/2$ , where  $\Delta\Pi$  is the change in osmotic pressure upon swelling and  $\varepsilon_d$  is the swelling strain. For magnetically actuated materials with programmed ferromagnetic domains, the energy density was approximated as  $\rho_E = 3\mu(\varepsilon_a)^2/2$ , where  $\mu$  is the shear modulus and  $\varepsilon_a$  is the strain that developed in a deformed state due to the applied magnetic field. The quantities to evaluate the actuation performance were summarized in Extended Data Fig. 8.

**Data availability.** All data generated or analysed during this study are included in the published article and its Supplementary Information, and are available from the corresponding author on reasonable request.

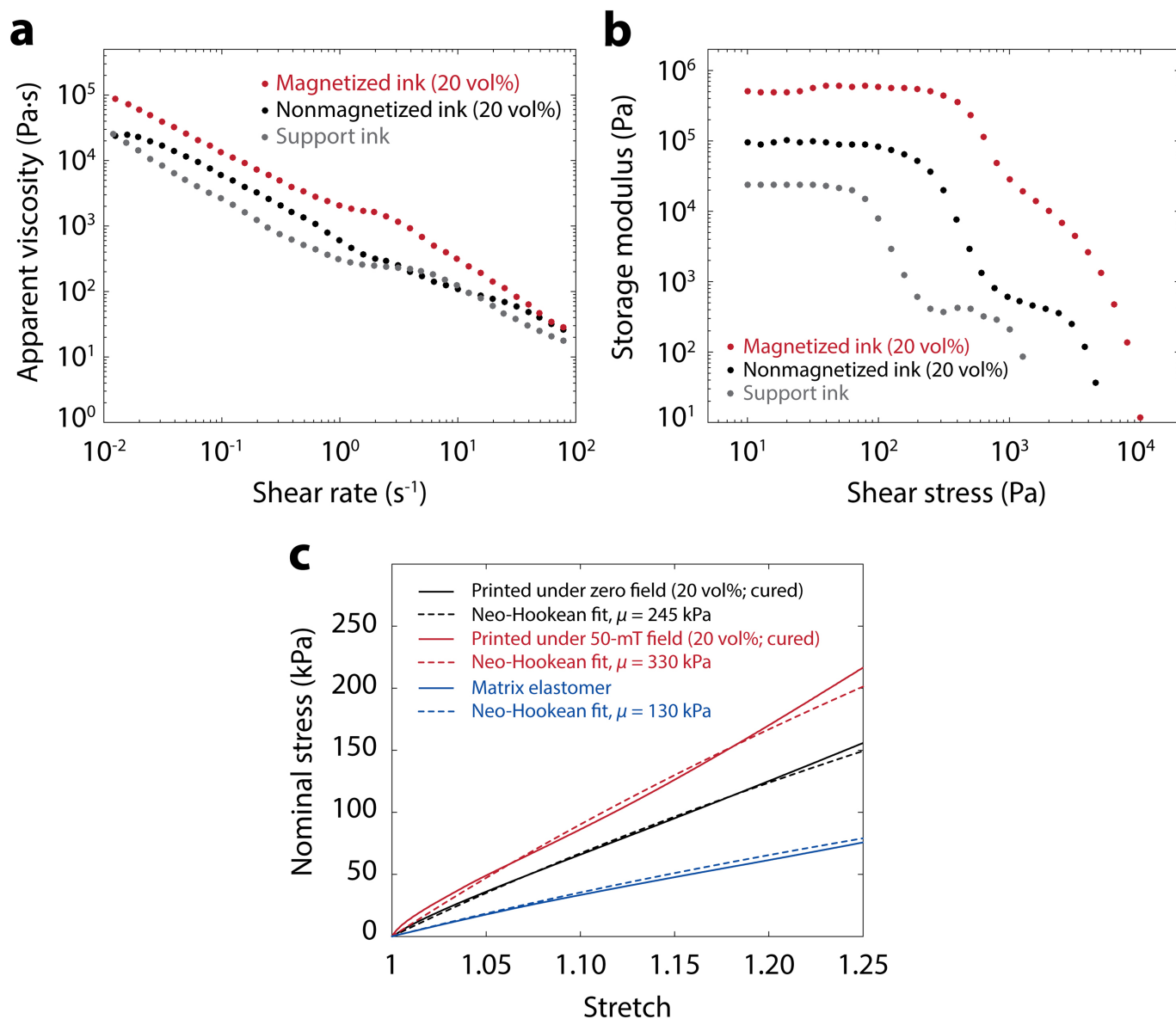
**Code availability.** The codes for 3D printing and the user element subroutine for numerical simulation are available upon request.

32. Zhang, Q., Zhang, K. & Hu, G. Smart three-dimensional lightweight structure triggered from a thin composite sheet via 3D printing technique. *Sci. Rep.* **6**, 22431 (2016).
33. Mirfakhrai, T., Madden, J. D. & Baughman, R. H. Polymer artificial muscles. *Mater. Today* **10**, 30–38 (2007).
34. Illeperuma, W. R., Sun, J.-Y., Suo, Z. & Vlassak, J. J. Force and stroke of a hydrogel actuator. *Soft Matter* **9**, 8504–8511 (2013).
35. Huang, L. M. et al. Ultrafast digital printing toward 4D shape changing materials. *Adv. Mater.* **29**, 1605390 (2017).
36. Bakarich, S. E., Gorkin, R., Panhuis, M. h. & Spinks, G. M. 4D printing with mechanically robust, thermally actuating hydrogels. *Macromol. Rapid Commun.* **36**, 1211–1217 (2015).
37. Wu, J. et al. Multi-shape active composites by 3D printing of digital shape memory polymers. *Sci. Rep.* **6**, 24224 (2016).
38. Ambulo, C. P. et al. Four-dimensional printing of liquid crystal elastomers. *ACS Appl. Mater. Interfaces* **9**, 37332–37339 (2017).
39. Kotikian, A., Truby, R. L., Boley, J. W., White, T. J. & Lewis, J. A. 3D printing of liquid crystal elastomeric actuators with spatially programmed nematic order. *Adv. Mater.* **30**, 1706164 (2018).
40. Li, W. et al. Flexible circuits and soft actuators by printing assembly of graphene. *ACS Appl. Mater. Interfaces* **8**, 12369–12376 (2016).



**Extended Data Fig. 1 | Scanning electron microscope images of NdFeB and fumed silica particles and printed fibres. a,** Magnetizable microparticles of NdFeB alloy in flake-like shapes with an average size of  $5\ \mu\text{m}$ . **b,** Fumed silica nanoparticles with an average size of  $30\ \text{nm}$ .

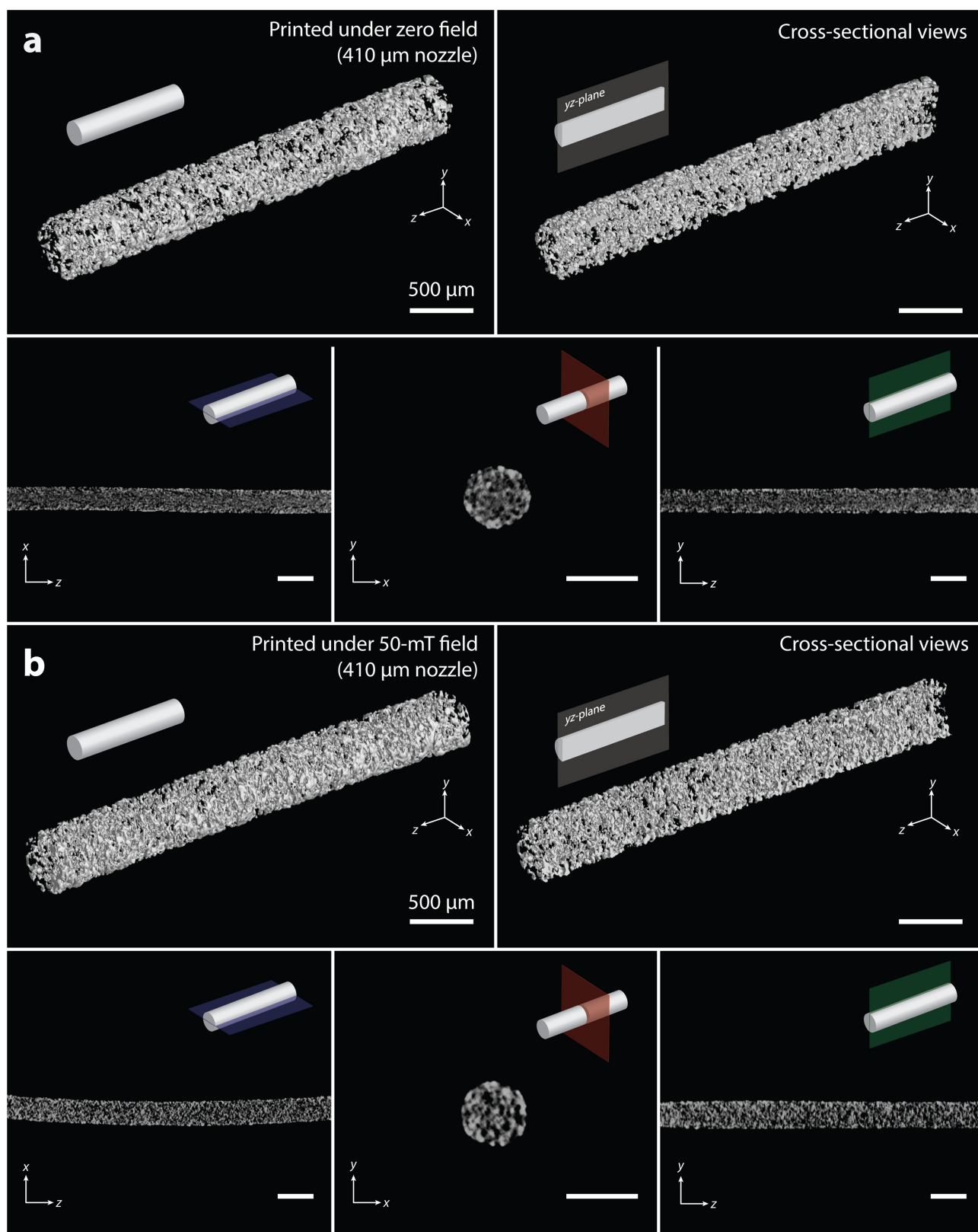
**c–f,** Fibres printed using nozzles with diameters of  $50\ \mu\text{m}$  (**c**),  $100\ \mu\text{m}$  (**d**),  $200\ \mu\text{m}$  (**e**), and  $410\ \mu\text{m}$  (**f**). **g,** The ratio between the printed fibre and the nozzle diameter, which is called the die-swelling ratio, plotted against the nozzle diameter.



**Extended Data Fig. 2 | Mechanical characterizations of the ink and the printed materials.** **a**, **b**, Apparent viscosity as a function of applied shear rate (**a**) and storage modulus as a function of applied shear stress (**b**) for 20 vol% magnetized ink (red), 20 vol% nonmagnetized ink (black) and support ink (grey). **c**, Nominal tensile stress–stretch curves (solid lines) for specimens printed with the magnetic ink in the absence of external

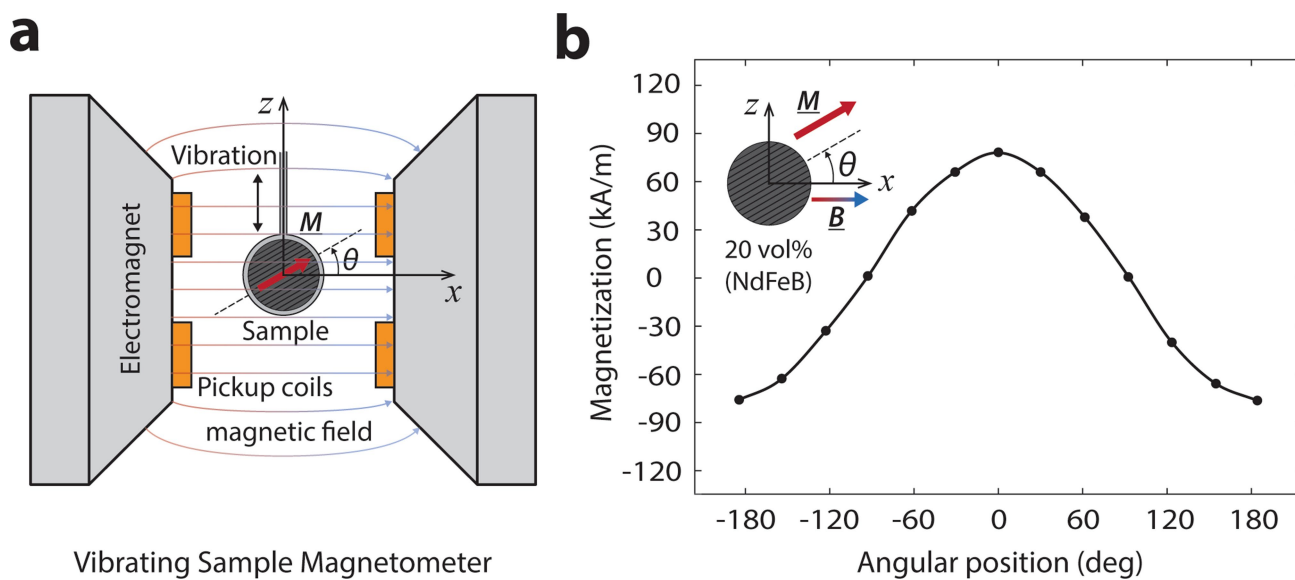
fields (black), with applied magnetic fields of 50 mT (red) at the nozzle tip generated by a permanent magnet, and the elastomer matrix with no magnetic particles (blue). The shear modulus  $\mu$  of each material was obtained by fitting the experimental curves to a neo-Hookean model (dashed lines).





**Extended Data Fig. 3 | Micro-computed tomography images of printed fibres.** **a**, A fibre printed with a nozzle of diameter 410  $\mu\text{m}$  in the absence of applied magnetic field. **b**, A fibre printed with a nozzle of diameter

410  $\mu\text{m}$  in the presence of an applied magnetic field of 50 mT at the nozzle tip. No obvious aggregation of ferromagnetic particles in the printed fibres can be observed.



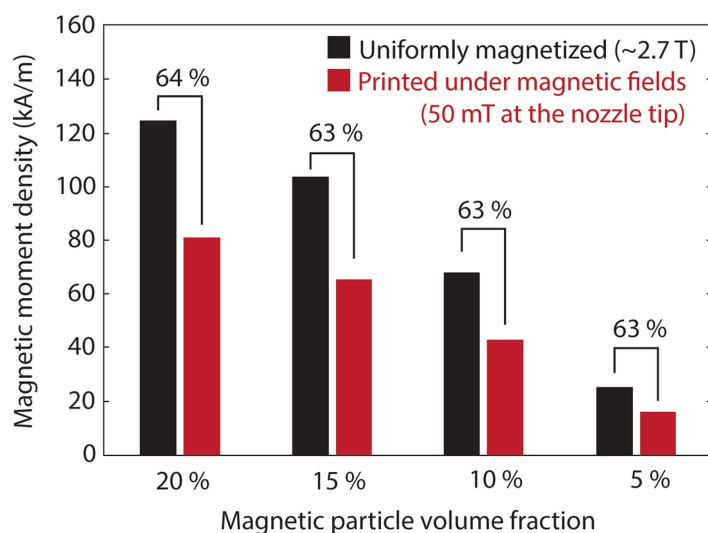
### Vibrating Sample Magnetometer

**Extended Data Fig. 4 | Experimental validation of the magnetization induced during printing under the applied magnetic field.**

**a**, Experimental setup with a vibrating sample magnetometer for measuring the magnetization of a sample printed with a nozzle of diameter

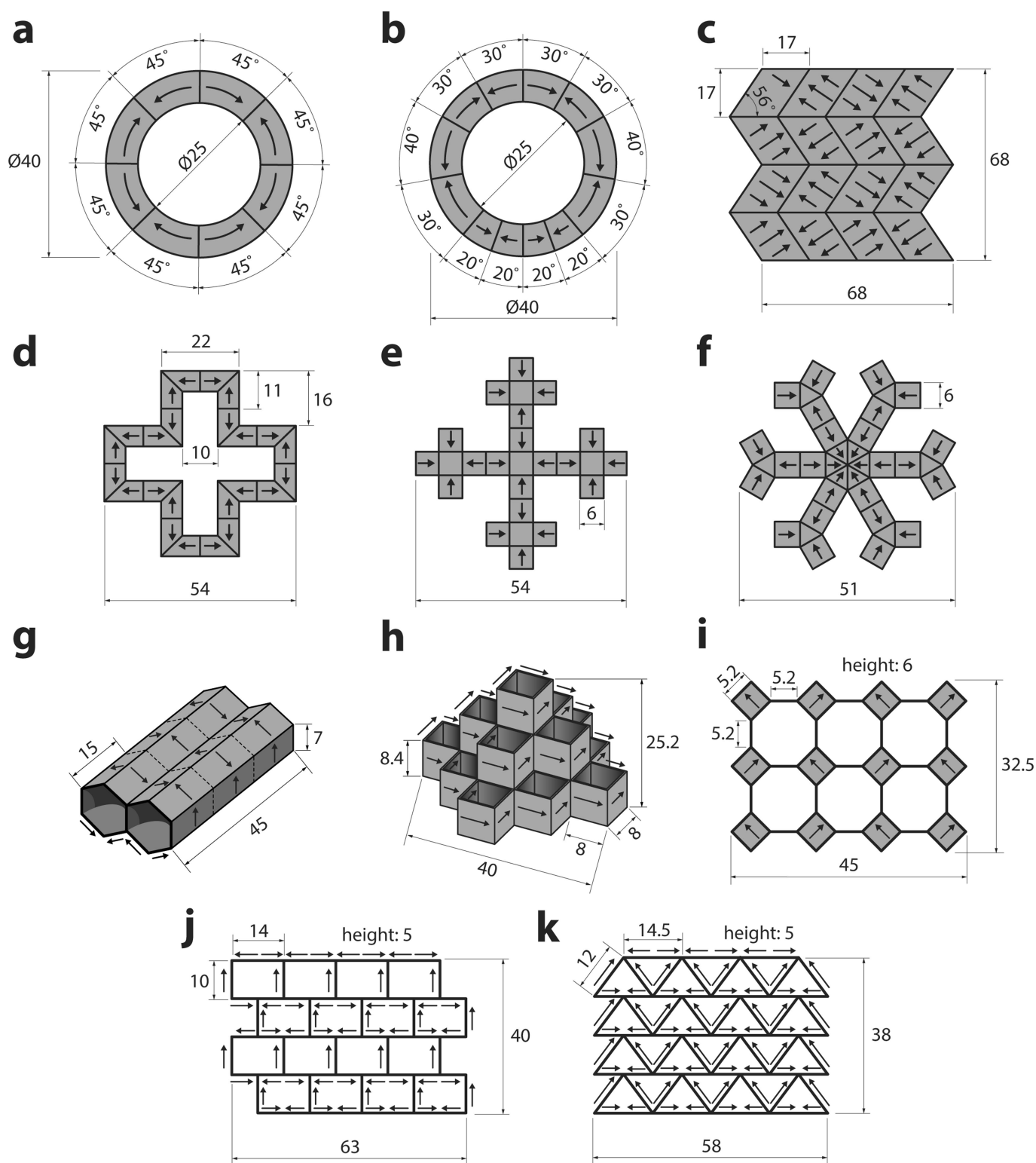
410  $\mu\text{m}$  under the presence of magnetic field of 50 mT at the nozzle tip.

**b**, Magnetization values measured at various angular positions of the printed fibres with respect to the external magnetic field applied by the vibrating sample magnetometer.



**Extended Data Fig. 5 | Magnetic moment densities of printed samples with different NdFeB particle volume fractions.** For printed specimens at each volume fraction, a permanent magnet is used to generate external fields (50 mT at the nozzle tip). The magnetization values of printed samples (red) are compared with the maximum achievable magnetization values (black) measured from uniformly magnetized samples. The

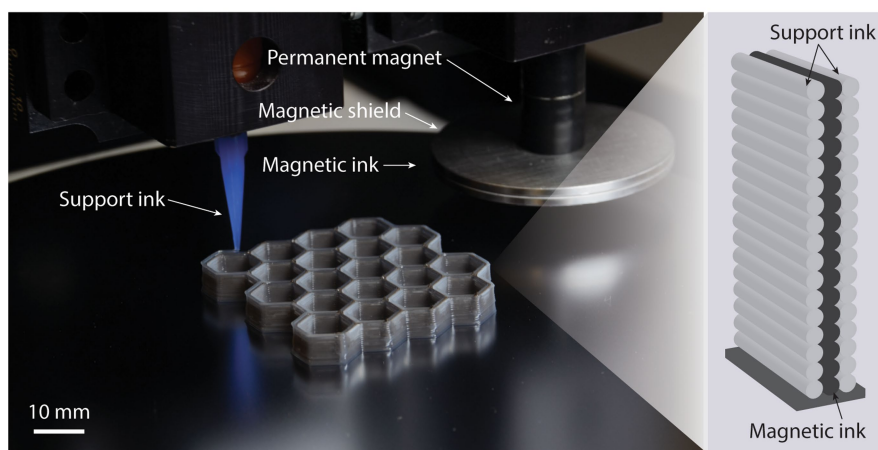
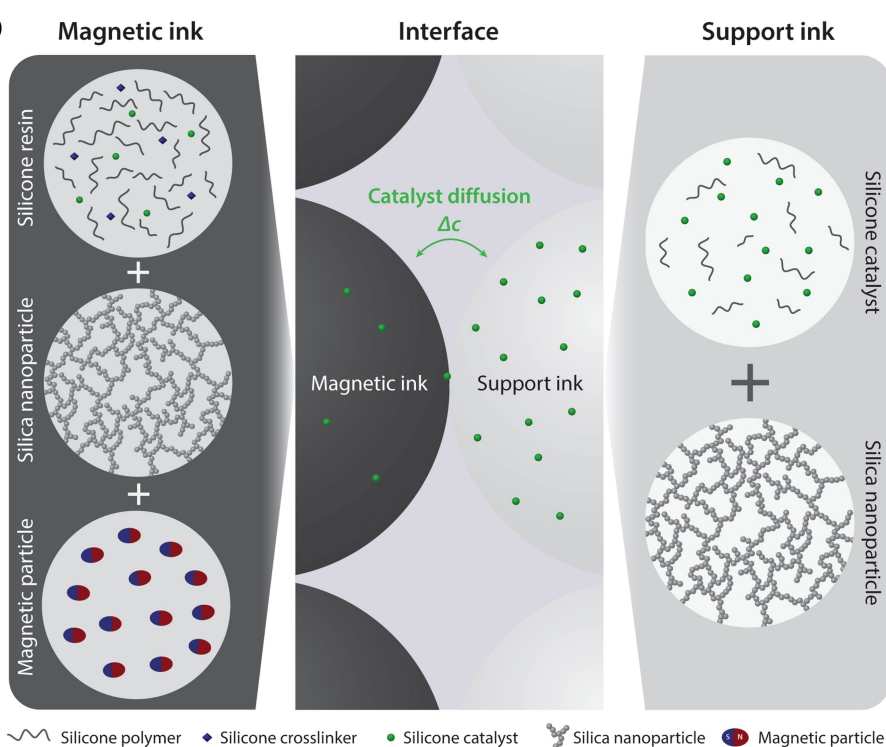
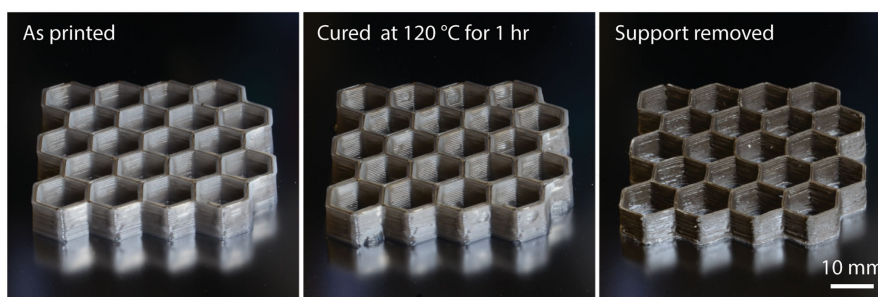
uniformly magnetized samples are printed in the absence of external fields, cured, and then magnetized under impulsive fields (about 2.7 T). Printing under the applied field of 50 mT yields a magnetic moment density that corresponds to 63%–64% of the maximum achievable magnetization at each volume fraction of NdFeB particles.



**Extended Data Fig. 6 | Schematic designs and dimensions of the two-dimensional and 3D structures in Figs. 2 and 3.** **a**, An annulus encoded with alternating domains that are equidistant. **b**, An annulus encoded with alternating domains that vary in size. **c**, A Miura-ori fold encoded with alternating oblique patterns of ferromagnetic domains. **d**, A hollow cross encoded with alternating ferromagnetic domains along the perimeter. **e**, **f**, Quadrupedal (**e**) and hexapedal (**f**) structures enabled by folding of the magnetically active segments surrounding the magnetically inactive

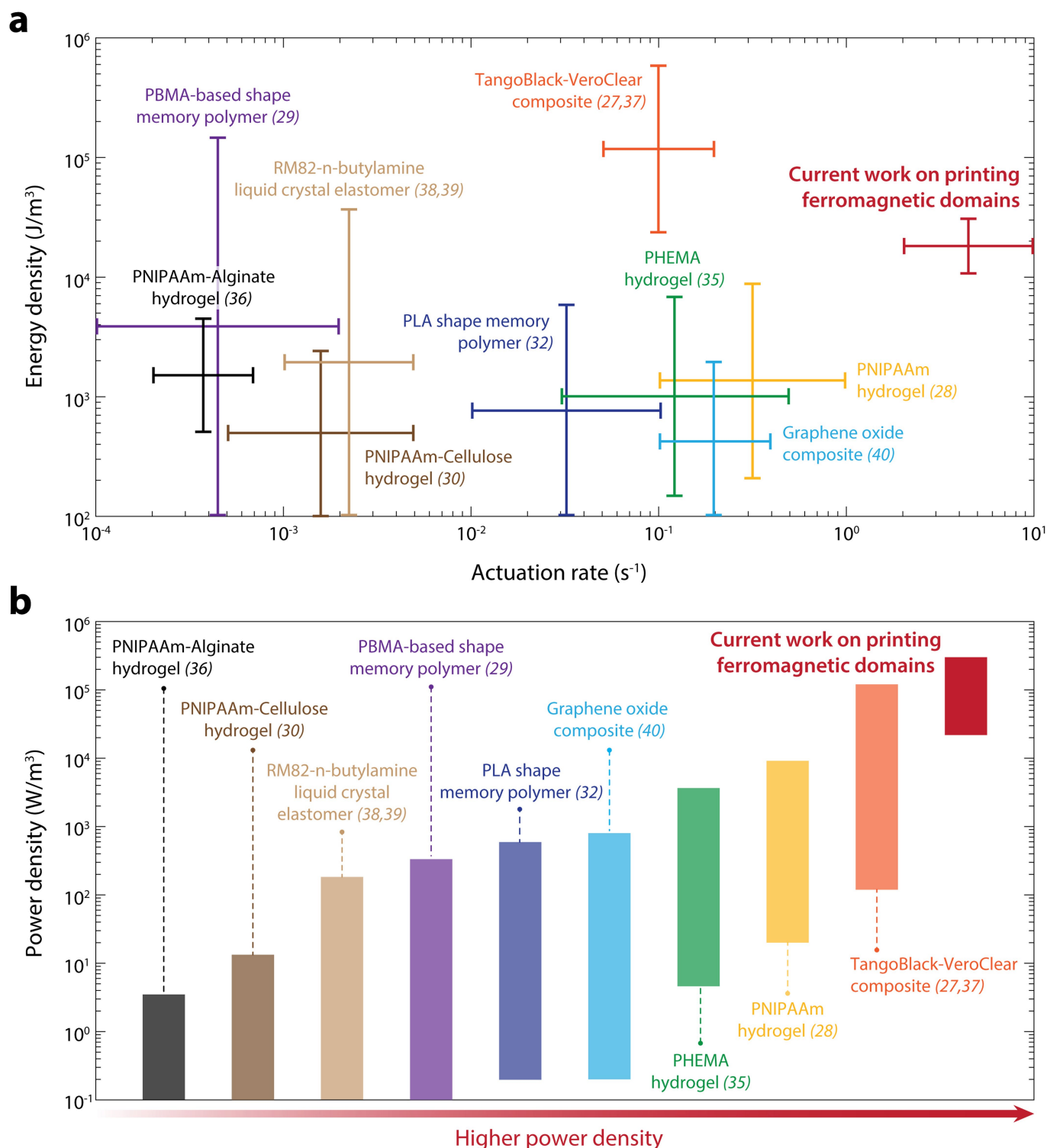
segments. **g**, Two adjoining hexagonal tubes programmed to form undulating surfaces under applied magnetic fields. **h**, A pyramid-shaped thin-walled structure programmed to elongate in its diagonal direction along applied magnetic fields. **i–k**, A set of auxetic structures (with negative Poisson's ratios) programmed to shrink in both length and width under applied magnetic fields. The dimensions in this figure are given in millimetres.



**a****b****c**

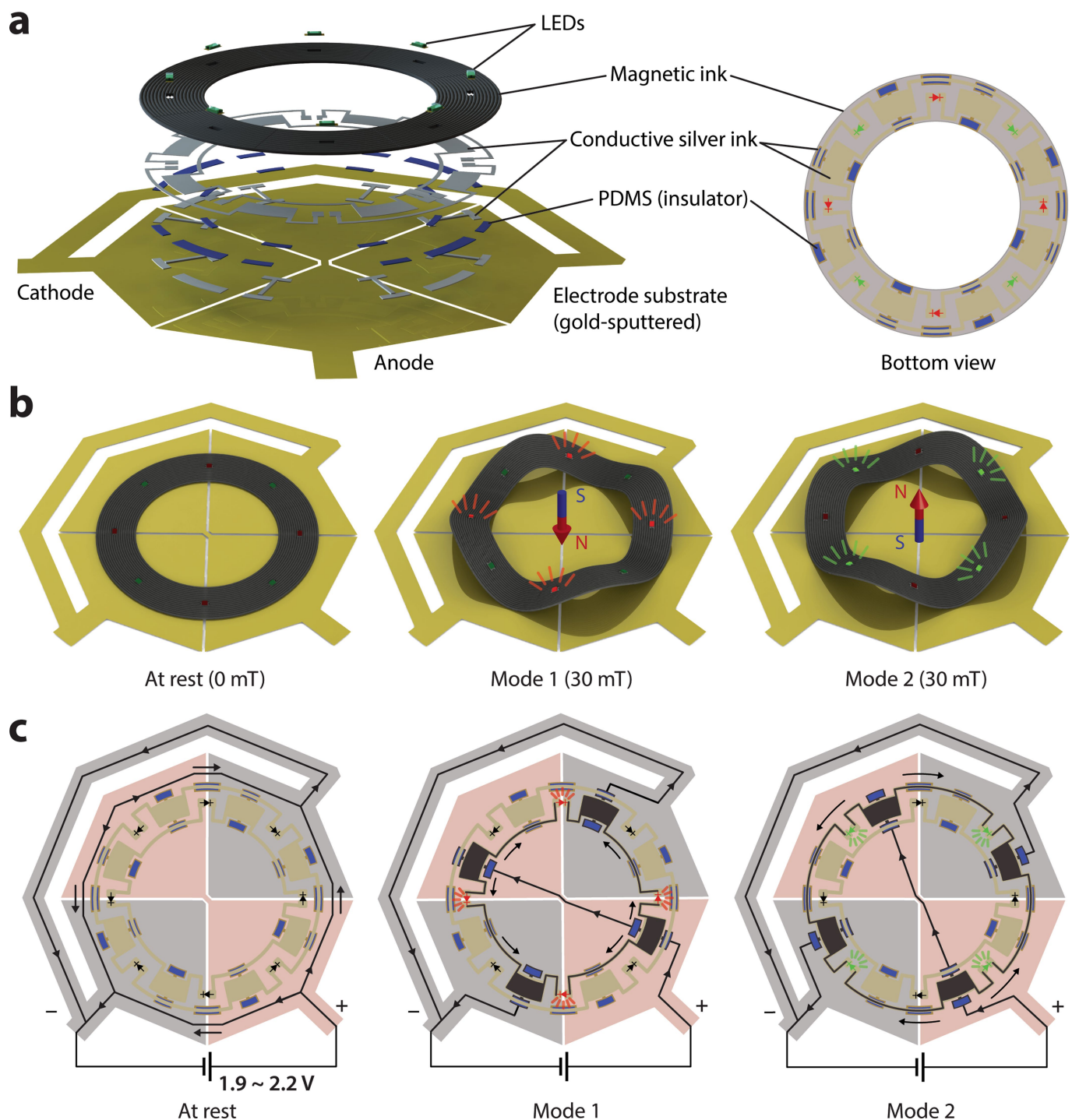
**Extended Data Fig. 7 | Overall fabrication process of printing multilayered structures assisted by the use of support inks. a,** Printing multilayered hexagonal arrays using magnetic and support inks. The use of support inks as fugitive buttresses enables stacking the deposited magnetic inks stably up to tens or even hundreds of layers. **b,** Chemical composition of magnetic and support inks. The higher concentration of catalyst in the

support ink prevents diffusion of catalyst molecules through the interface and thus prevents imperfect curing of the adjacent magnetic inks.  $\Delta c$  denotes the difference in catalyst concentration between the support and magnetic inks. **c,** The printed magnetic inks are cured by heating at 120 °C for 1 h. The support ink is then removed by solvent rinses.



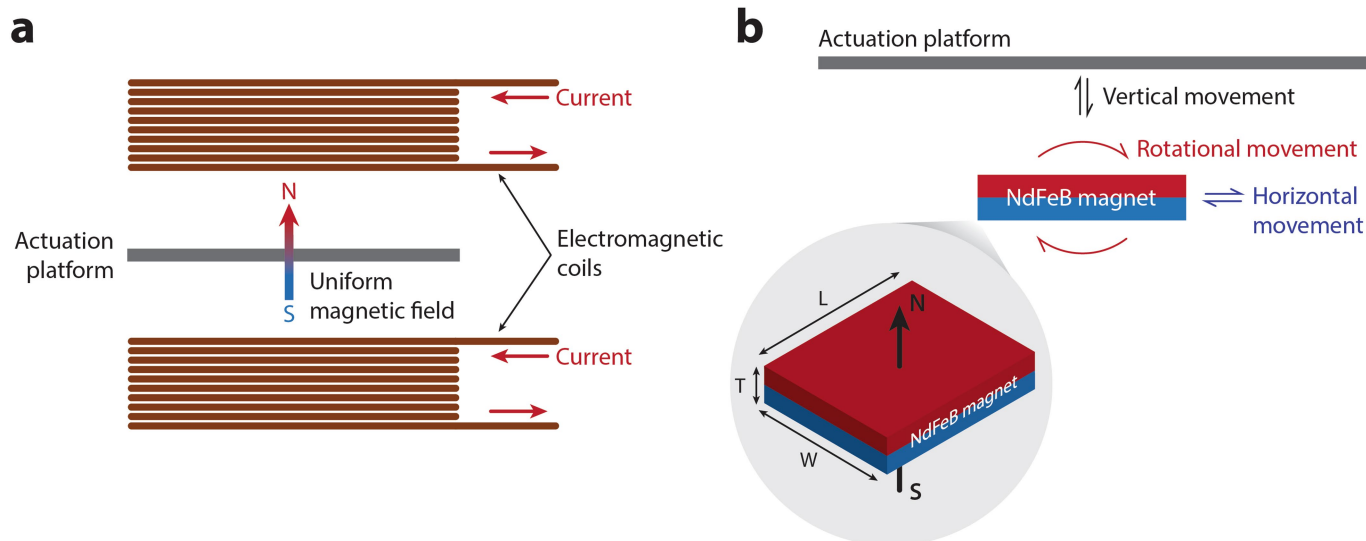
**Extended Data Fig. 8 | Actuation performance of 3D-printed shape-programmable soft materials. a,** Energy density and actuation rate of our magnetically responsive structures presented in Figs. 1–3 are plotted and compared with those of existing 3D-printed shape-programmable soft materials in the literature<sup>27–30,32,35–40</sup>. **b,** Power density is calculated as energy density multiplied by the actuation rate of each material and plotted for comparison; the materials are listed in order of

increasing power densities. PNIPAAm = poly(*N*-isopropylacrylamide); PMBA = poly(benzyl methacrylate); PHEMA = poly(hydroxyl ethyl methacrylate); PLA = poly(lactic acid); RM82 = 1,4-bis-[4-(6-acryloyloxyhexyloxy)benzoyloxy]-2-methylbenzene. TangoBlack and VeroClear are commercially available acrylic photocurable polymers from Stratasys Ltd.



**Extended Data Fig. 9 | Schematic designs and working principles of the reconfigurable soft electronic device demonstrated in Fig. 4a.** **a**, Exploded and bottom views of the printed device, in which soft electronic circuitry and components are embedded by means of a hybrid fabrication process based on multimaterial 3D printing. **b**, Two different shapes depending on the direction of applied magnetic fields of 30 mT,

which yield different electronic functions (red micro-LEDs lit up in Mode 1 and green micro-LEDs lit up in Mode 2). **c**, Schematic diagram of the embedded soft electronic circuits, which are designed to turn active only in the designated mode of transformation owing to the selective contact with the gold electrode on the substrate.



**Extended Data Fig. 10 | Schematic illustrations of the methods for applying magnetic fields to actuate the printed structures. a, b,** The magnetic fields for actuating the printed structures can be applied in two ways. A pair of electromagnetic coils are used to generate a uniform

magnetic field (a). A NdFeB magnet (width 2 in, length 3 in, thickness 0.5 in, surface flux density 300 mT) is used to create spatially varying magnetic fields for dynamic actuation by combining vertical, horizontal and rotational movements of the magnet (b).



# Velocity-resolved kinetics of site-specific carbon monoxide oxidation on platinum surfaces

Jannis Neugebohren<sup>1</sup>, Dmitriy Borodin<sup>1</sup>, Hinrich W. Hahn<sup>1</sup>, Jan Altschäffell<sup>1,2</sup>, Alexander Kandratsenka<sup>2</sup>, Daniel J. Auerbach<sup>2</sup>, Charles T. Campbell<sup>3</sup>, Dirk Schwarzer<sup>2</sup>, Dan J. Harding<sup>1,2,7</sup>, Alec M. Wodtke<sup>1,2,4</sup> & Theofanis N. Kitsopoulos<sup>2,5,6\*</sup>

**Catalysts are widely used to increase reaction rates. They function by stabilizing the transition state of the reaction at their active site, where the atomic arrangement ensures favourable interactions<sup>1</sup>. However, mechanistic understanding is often limited when catalysts possess multiple active sites—such as sites associated with either the step edges or the close-packed terraces of inorganic nanoparticles<sup>2–4</sup>—with distinct activities that cannot be measured simultaneously. An example is the oxidation of carbon monoxide over platinum surfaces, one of the oldest and best studied heterogeneous reactions. In 1824, this reaction was recognized to be crucial for the function of the Davy safety lamp, and today it is used to optimize combustion, hydrogen production and fuel-cell operation<sup>5,6</sup>. The carbon dioxide products are formed in a bimodal kinetic energy distribution<sup>7–13</sup>; however, despite extensive study<sup>5</sup>, it remains unclear whether this reflects the involvement of more than one reaction mechanism occurring at multiple active sites<sup>12,13</sup>. Here we show that the reaction rates at different active sites can be measured simultaneously, using molecular beams to controllably introduce reactants and slice ion imaging<sup>14,15</sup> to map the velocity vectors of the product molecules, which reflect the symmetry and the orientation of the active site<sup>16</sup>. We use this velocity-resolved kinetics approach to map the oxidation rates of carbon monoxide at step edges and terrace sites on platinum surfaces, and find that the reaction proceeds through two distinct channels<sup>11–13</sup>: it is dominated at low temperatures by the more active step sites, and at high temperatures by the more abundant terrace sites. We expect our approach to be applicable to a wide range of heterogeneous reactions and to provide improved mechanistic understanding of the contribution of different active sites, which should be useful in the design of improved catalysts.**

The oxidation of CO on the (111) surface of platinum—the simplest close-packed single-crystal surface—proceeds through the Langmuir–Hinshelwood mechanism<sup>7,17</sup>. O<sub>2</sub> is activated by dissociating into adsorbed oxygen atoms, which combine with adsorbed CO to form CO<sub>2</sub> with substantial excess energy. This can lead to hyperthermal CO<sub>2</sub>, which has high kinetic and internal energy<sup>8,9</sup> and exhibits angular distributions that are narrower than those expected for a molecule at thermal equilibrium with the surface<sup>7,9,10</sup>; however, some reaction conditions also result in a thermal channel<sup>11–13</sup>. Although this bimodal generation of products suggests that CO oxidation occurs through more than one reaction<sup>12,13</sup>, the dominant view at present is that all products form through the same transition state, and a fraction desorbs promptly with hyperthermal velocities whereas the rest become trapped and thermalized at the surface before desorbing<sup>11</sup>.

Our study of this reaction exploits the correlation between the angle and velocity of reaction products and the symmetry and orientation of surface reaction sites<sup>16</sup>. This enabled us to extract site-specific kinetic data by data-mapping the temporal evolution of velocity-resolved products. We used pulsed-molecular-beam surface dosing of both CO

and O<sub>2</sub> in combination with slice ion imaging<sup>14,15</sup>. The O<sub>2</sub> beam was used to control oxygen coverage and the pulsed CO beam was used to initiate the reaction; together with synchronized pulsed-laser ionization of CO<sub>2</sub>, this enabled us to obtain the CO<sub>2</sub> flux as a function of reaction time. At each time point in the reaction, ion imaging provided distributions of the velocity vectors of the products.

For CO oxidation on Pt(111) with a step density of 0.25%, ion images of the CO<sub>2</sub> produced (Fig. 1a) show a bimodal speed distribution<sup>12,13</sup> (Fig. 1b). The low-velocity component was fit to a Maxwell–Boltzmann distribution, whereas the high-velocity component is hyperthermal. The angular distributions of the two channels (Fig. 1c) show that the hyperthermal component has a sharp peak along the surface normal (approximately  $\cos^8\theta$ ), whereas the thermal channel follows a  $\cos\theta$  distribution.

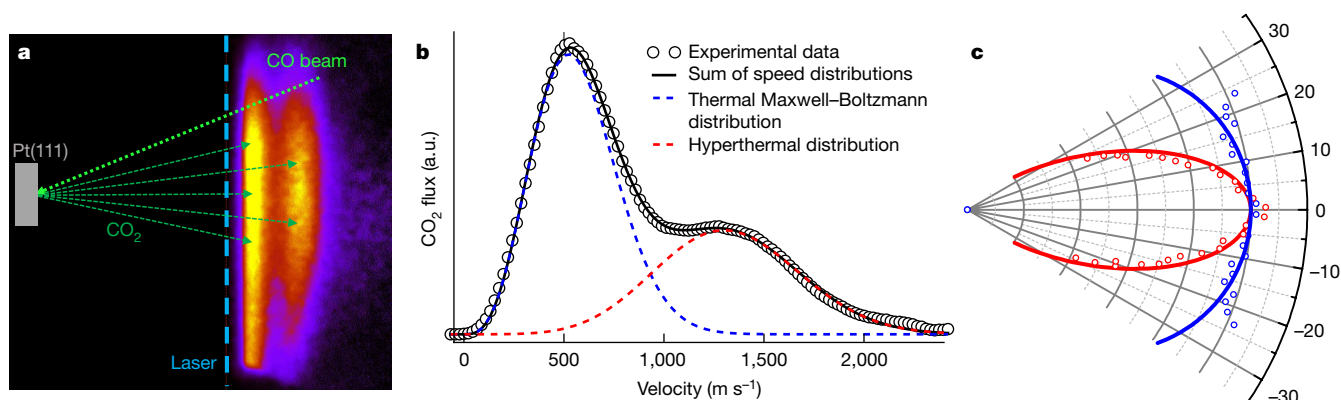
To obtain site-specific kinetics, we recorded the integral of the ion image intensity over selected velocity ranges (red and blue rectangles in the inset of Fig. 2) as a function of the delay between the CO and laser pulses. The raw time-dependent signals are proportional to product density and depend not only on the reaction time at the surface, but also on the flight time of the CO<sub>2</sub> from the surface to the ionizing laser spot. By recording the velocity of the products (see Methods), we corrected for the CO<sub>2</sub> flight-time and obtained the relative flux in both channels. From the measured angular distributions, we corrected the signal strength to reflect the true branching ratio between the two channels. Thus, we rigorously obtained the relative product flux as a function of reaction time, hereafter referred to as the kinetic trace.

Kinetic traces are shown in Fig. 2 for both the hyperthermal (red) and thermal (blue) channels. In both, the CO<sub>2</sub> formation rate first increases as CO adsorbs onto the surface (black dashed line), and then decays with an effective lifetime,  $\tau_{\text{eff}}$  (see Methods), as the adsorbed CO is removed.  $\tau_{\text{eff}}$  for the thermal and hyperthermal channels—hereafter referred to as  $\tau_{\text{eff}}^{\text{Step}}$  and  $\tau_{\text{eff}}^{\text{Terr}}$ , respectively—are quite different;  $\tau_{\text{eff}}^{\text{Terr}}$  is shorter than  $\tau_{\text{eff}}^{\text{Step}}$  under all conditions in this study. This is inconsistent with a bimodal velocity distribution resulting from partial thermalization of products from a single reaction<sup>7,11</sup>, as that would require the two kinetic traces to be identical. It should be noted that the desorption of physisorbed CO<sub>2</sub> is fast compared to the reaction time.

We assign the hyperthermal and thermal channels to the reaction of CO with oxygen adatoms at terraces and steps, respectively, for three reasons. First, when we increased the step density 66-fold using a Pt(332) crystal, the hyperthermal channel disappeared completely at low oxygen coverage, at which nearly all oxygen atoms are bound at steps (lower left inset of Fig. 2). This confirms that the hyperthermal channel does not occur at steps. Second, in the limit of zero oxygen coverage,  $\tau_{\text{eff}}$  is determined by any non-reactive channels that consume adsorbed CO (see Methods). For the thermal channel, the zero-coverage rate constants (filled blue circles in Extended Data Fig. 4) are in excellent agreement with known values of CO desorption from steps<sup>18</sup>, confirming that CO molecules adsorbed at steps (hereafter

<sup>1</sup>Institute for Physical Chemistry, University of Goettingen, Göttingen, Germany. <sup>2</sup>Department of Dynamics at Surfaces, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

<sup>3</sup>Department of Chemistry, University of Washington, Seattle, WA, USA. <sup>4</sup>International Center for Advanced Studies of Energy Conversion, Georg-August University of Göttingen, Göttingen, Germany. <sup>5</sup>Department of Chemistry, University of Crete, Heraklion, Greece. <sup>6</sup>Institute of Electronic Structure and Laser, FORTH, Heraklion, Greece. <sup>7</sup>Present address: Department of Chemical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. \*e-mail: tkitsop@mpibpc.mpg.de



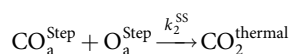
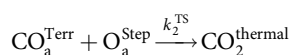
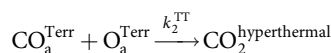
**Fig. 1 | Slice ion imaging of CO<sub>2</sub> produced by CO oxidation at Pt(111).** **a**, Ion image of CO<sub>2</sub> produced by the catalytic oxidation of CO on a Pt(111) surface. The velocity vector of the incident CO beam is shown by the light green arrow, as are a few of the CO<sub>2</sub> product vectors appearing in the thermal and hyperthermal channels. CO<sub>2</sub> speed increases to the right. The position of the ionizing laser beam is shown by the dashed blue line. **b**, Velocity probability distribution of CO<sub>2</sub> molecules. The thermal

product channel is fitted to a Maxwell–Boltzmann function (translational temperature,  $T_{\text{trans}} = 483$  K; dashed blue line) and the hyperthermal channel is fit to a flowing Maxwell–Boltzmann function ( $T_{\text{trans}} = 894$  K,  $\alpha = 190$  meV; dashed red line). **c**, Velocity-resolved angular distributions: hyperthermal (red) and thermal (blue) channels. The hyperthermal channel is compared to a  $\cos^6\theta$  function (red line) and the thermal channel is compared to a  $\cos\theta$  function (blue line). a.u., arbitrary units.

denoted  $\text{CO}_a^{\text{Step}}$  are important to the thermal reaction. Third, we calculated the minimum energy path of the terrace reaction  $\text{CO}_a^{\text{Terr}} + \text{O}_a^{\text{Terr}}$  using density functional theory (DFT; see Methods). This shows an energy release of 1.9 eV and an early transition-state structure (that is, it resembles the reactants), which is consistent with the average translational energy release of 0.38 eV that is seen in the hyperthermal channel. This assignment is also consistent with the previously observed dominance of the hyperthermal channel at 880 K<sup>8</sup> (see Methods) and the prevalence of the thermal channel that is seen with increased step density<sup>11</sup>.

Considering the above, we interpret the kinetic traces of Fig. 2 in terms of competing reactions that occur with oxygen bound at terrace and step sites. We consider the terrace reaction competing with a process involving diffusion that converts  $\text{CO}_a^{\text{Terr}}$  to  $\text{CO}_a^{\text{Step}}$ . This explains how adsorption at terraces (more than 99% of the CO adsorbs initially at terraces, as the Pt(111) step density is less than 1%) can lead to more CO<sub>2</sub> production at steps. The kinetic trace of the thermal channel shown in Fig. 2 cannot be fit with a single exponential function (see Extended Data Figs. 6, 8), which suggests the presence of two step reactions.

In Methods, we present a simple kinetic model that describes our data over the full range of oxygen coverages and surface temperatures using three reactions:

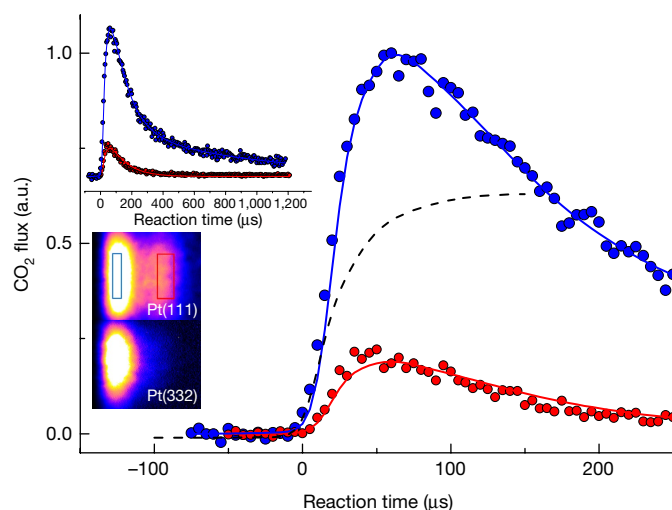


We also include  $\text{CO}_a^{\text{Step}}$  desorption, as well as  $\text{CO}_a^{\text{Terr}}$  desorption and conversion of  $\text{CO}_a^{\text{Terr}}$  to  $\text{CO}_a^{\text{Step}}$ . The determination of oxygen coverage and the  $\text{O}_a^{\text{Step}}/\text{O}_a^{\text{Terr}}$  ratio is described in Methods. We note that velocity-resolved kinetics of CO oxidation on Pt(332), where the step density is 66-fold higher, supports the three-reaction mechanism.

The derived pre-exponential factors and activation energies are shown in Table 1 and the temperature dependent reaction rate constants are shown in Fig. 3. The results obtained on the Pt(111) and the Pt(332) crystals agree well. The rate constants for CO oxidation at steps are similar to those for reaction at terraces. Despite this, the product yield through step reactions dominates at all temperatures used in this work and in previous studies (area to the left of the vertical line in Fig. 3). Even for a Pt(111) crystal with a step density of only 0.25%, reactions at steps dominate. Analysis with our kinetic model shows that

this reflects the fact that conversion of  $\text{CO}_a^{\text{Terr}}$  to  $\text{CO}_a^{\text{Step}}$  is much faster than the TT reaction in the temperature range of this and previous work. We therefore conclude that all previous kinetics studies were strongly influenced by step reactions (see Methods).

For this reason, this is the first report to our knowledge of the rate parameters for the Pt(111) terrace reaction,  $k_2^{\text{TT}}$ . Only the rate of the TT reaction has been calculated from first principles<sup>19,20</sup>; as such, the new rate parameters provide the first opportunity to benchmark the ab



**Fig. 2 | Kinetic traces of CO<sub>2</sub> obtained from velocity-resolved kinetics experiments.** The inset labelled Pt(111) shows an ion image with velocity-space integration windows for the hyperthermal (red; 1,280–1,610 m s<sup>−1</sup>) and the thermal (blue; 420–590 m s<sup>−1</sup>) channels. The dashed black line represents the measured CO dosing function; the onset of the incident CO beam is taken as the zero of time. The solid red and blue lines are the fits resulting from the kinetic model (see Methods). The two traces are normalized to reflect the experimental time-integrated branching ratio between the two channels (8.6:1 in favour of the thermal channel) at this temperature and oxygen coverage. The step density was 0.0025 monolayers, the oxygen coverage was 0.04 monolayers and the dose of a single CO pulse was  $2 \times 10^{-5}$  monolayers. The surface temperature was 593 K. The experimental conditions for the (332) image are: oxygen coverage, 0.04 monolayers; surface temperature, 593 K; step density, 0.167 monolayers. Here, oxygen atoms are expected to be entirely at step sites, because O<sub>2</sub> dissociates more easily at steps than at terraces<sup>29</sup> and oxygen atoms are bound more strongly at steps than at terraces<sup>30</sup>. For the reaction on the (332) sample, the hyperthermal channel is absent.

**Table 1 | Rate parameters for the elementary processes involved in the site-specific oxidation of carbon monoxide on a platinum catalyst**

	Elementary process	$A$ ( $s^{-1}$ )	$E_a$ (eV)
$k_0^T$	Desorption from terraces <sup>a</sup>	$5.9^{+5.4}_{-2.8} \times 10^{13}$	$1.28 \pm 0.02$
$k_0^S$	Desorption from steps <sup>b</sup>	$1.5^{+2.5}_{-0.6} \times 10^{12}$	$1.18 \pm 0.056$
$k_1^{TS}$ 111 (332)	Diffusion from terrace to step <sup>c</sup>	$2.1^{+13}_{-1.8} \times 10^6$ ( $1.5 \times 10^6$ )	$0.30 \pm 0.1$ ( $0.32 \pm 0.1$ )
$k_2^{TT}$ 111 (332)	$CO_{Terr} + O_{Terr}$ reaction	$3.5^{+21}_{-3.0} \times 10^9$ ( $3.8^{+23}_{-3.3} \times 10^9$ )	$0.6 \pm 0.1$ ( $0.6 \pm 0.1$ )
<b>Theory<sup>19</sup></b>		<b><math>5.1 \times 10^{12}</math></b>	<b>0.74</b>
$k_2^{TS}$ 111 (332)	$CO_{Terr} + O_{Step}$ reaction	$5.9^{+36}_{-5.1} \times 10^7$ ( $4.1^{+25}_{-3.5} \times 10^7$ )	$0.40 \pm 0.1$ ( $0.40 \pm 0.1$ )
$k_2^{SS}$ 111 (332)	$CO_{Step} + O_{Step}$ reaction	$2.9^{+18}_{-2.5} \times 10^9$ ( $2.6^{+16}_{-2.2} \times 10^9$ )	$0.65 \pm 0.1$ ( $0.65 \pm 0.1$ )

The rate parameters reported here should be used with caution outside the temperature range used in this work.

<sup>a</sup>Measured independently, as published in ref. 26. This value also compares well with previous reports. See ref. 27:  $A = 3.5^{+7.2}_{-2.7} \times 10^{13} s^{-1}$ ,  $E_a = 1.27 \pm 0.07$  eV.

<sup>b</sup>Obtained from measurements of desorption of CO from Pt(332). Compares well with rate constants reported in ref. 27.

<sup>c</sup>An activation energy of 0.3 eV was chosen following ref. 28. This should be considered an effective conversion rate.

initio rates of this important catalytic reaction. Figure 3 shows one such calculation of the rate constants of the TT reaction, for which DFT is used to generate input for a transition-state-theory calculation of the rate constant<sup>19</sup>. The theoretical rate constant is 20–100 times larger than that seen experimentally. The theoretically calculated activation energy is higher than that derived from experiment (Table 1), as is expected when using a revised Perdew–Burke–Ernzerhof (RPBE) functional<sup>21</sup>. Improving the calculation would lower the barrier; however, this would only amplify the disagreement with experiment. Thus, it is clear that the theory fails to reproduce the prefactor of the rate constant. This can be the result of an incorrect treatment of reactant entropy<sup>22</sup> and/or dynamical recrossing at the transition state.

Velocity-resolved kinetics provides a powerful tool for catalytic kinetic studies. Previous studies have attempted to obtain CO oxidation kinetics without velocity resolution by measuring phase shifts with modulated reactant molecular beams<sup>7,23</sup>. To analyse these phase-lag data, the kinetic trace was assumed to follow a single exponential decay<sup>7,23</sup>; this was effectively a single-reaction assumption that resulted

in an activation energy that was dependent on oxygen coverage<sup>7</sup>. However, we show in Methods that these were only apparent activation energies that result from the temperature and oxygen-coverage-dependent influence of the step and terrace reactions.

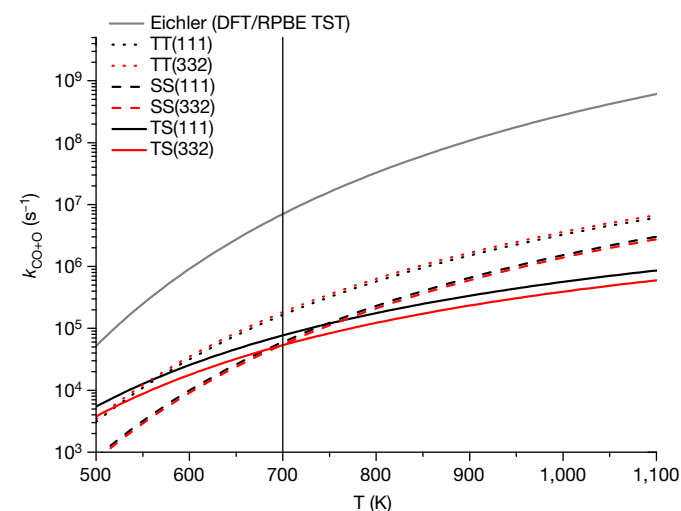
The phase-lag method is analogous to frequency domain kinetic methods that are used in biophysics for obtaining (exponential) fluorescence lifetimes<sup>24</sup>, and was the main method used before the advent of pulsed-laser-based pump–probe methods. In providing the kinetic trace directly, velocity-resolved kinetics serves the function of a pump–probe experiment for neutral matter. This enabled us to easily identify three elementary reactions that comprise a site-specific mechanism of CO oxidation on platinum. The activation energies for these three reactions are not dependent on oxygen coverage.

The relative importance of different reacting geometries occurring at different active sites has been described previously in terms of a geometry probability-weighted activation energy<sup>25</sup>:  $E_{ai}^w = E_{ai} - kT \ln A_i$ . Here  $A_i$  is the abundance of the  $i$ th reacting geometry and  $E_{ai}$  is its activation energy. The results of this work reveal that the values of  $A_i$  depend on diffusion and desorption, an outcome that is also relevant to real-world catalysts that are often operated at high temperatures. In Extended Data Fig. 10 we show that the terrace reaction dominates on Pt(111) and Pt(332) at high temperatures, whereas at low temperatures the step reactions dominate. At high temperature, CO desorption becomes much faster than CO diffusion; as such, the site of CO adsorption determines which reaction occurs. Because step edges are minority sites on both samples, adsorption (and therefore reaction) occurs primarily at terraces. Extended Data Fig. 10 also shows that, under high-temperature conditions in which CO desorption suppresses the importance of diffusion, the efficiency of the conversion from CO to CO<sub>2</sub> is also reduced. We suggest that this may be a general phenomenon: low-concentration, high-reactivity defects on surfaces will be important in low-temperature catalysis in which reactant diffusion is faster than desorption, whereas at high temperatures, the majority facet dominates the chemistry and the efficiency of the reaction decreases.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0188-x>.

Received: 21 September 2017; Accepted: 16 April 2018;  
Published online 13 June 2018.



**Fig. 3 | Rate constants for CO oxidation on platinum steps and terraces.** Three reactions are identified in this work: the reaction of  $CO_a^{Terr}$  with terrace-bound oxygen atoms (TT) and two reactions of step-bound oxygen atoms with  $CO_a^{Steps}$  (SS) and  $CO_a^{TS}$  (TS). Results obtained from a Pt(111) surface with a step density of 0.25% and those obtained from a Pt(332) crystal with a step density of 16.7% are shown as black and red lines, respectively. The units of the rate constant are those used when oxygen coverage is given as a (unitless) fraction of a monolayer. The grey line shows the results of transition-state theory (TST) calculations of the TT reaction rate based on a DFT calculation of the transition state<sup>19</sup>. It considerably overestimates reactivity. The region to the left of the vertical line shows the range of temperatures studied here and in previous studies. The region to the right of this line shows the range of temperatures typical for industrial catalysis.

- Garcia-Viloca, M., Gao, J., Karplus, M. & Truhlar, D. G. How enzymes work: analysis by modern rate theory and computer simulations. *Science* **303**, 186–195 (2004).
- Mavrikakis, M., Stoltze, P. & Nørskov, J. K. Making gold less noble. *Catal. Lett.* **64**, 101–106 (2000).
- Jaramillo, T. F. et al. Identification of active edge sites for electrochemical H<sub>2</sub> evolution from MoS<sub>2</sub> nanocatalysts. *Science* **317**, 100–102 (2007).
- Van Santen, R. A. Complementary structure sensitive and insensitive catalytic relationships. *Acc. Chem. Res.* **42**, 57–66 (2009).

5. Santra, A. K. & Goodman, D. W. Catalytic oxidation of CO by platinum group metals: from ultrahigh vacuum to elevated pressures. *Electrochim. Acta* **47**, 3595–3609 (2002).
6. Park, E. D., Lee, D. & Lee, H. C. Recent progress in selective CO removal in a H<sub>2</sub>-rich stream. *Catal. Today* **139**, 280–290 (2009).
7. Campbell, C. T., Ertl, G., Kuipers, H. & Segner, J. A molecular-beam study of the catalytic-oxidation of CO on a Pt(111) surface. *J. Chem. Phys.* **73**, 5862–5873 (1980).
8. Becker, C. A., Cowin, J. P., Wharton, L. & Auerbach, D. J. CO<sub>2</sub> product velocity distributions for CO oxidation on platinum. *J. Chem. Phys.* **67**, 3394 (1977).
9. Cao, G. Y., Moula, M. G., Ohno, Y. & Matsushima, T. Dynamics on individual reaction sites in steady-state carbon monoxide oxidation on stepped platinum(113). *J. Phys. Chem. B* **103**, 3235–3241 (1999).
10. Palmer, R. L. & Smith, J. N. J. Molecular beam study of CO oxidation on a (111) platinum surface. *J. Chem. Phys.* **60**, 1453 (1974).
11. Segner, J., Campbell, C. T., Doyen, G. & Ertl, G. Catalytic-oxidation of CO on Pt(111) - the influence of surface-defects and composition on the reaction dynamics. *Surf. Sci.* **138**, 505–523 (1984).
12. Poehlmann, E., Schmitt, M., Hoinkes, H. & Wilsch, H. Bimodal angular and velocity distributions of CO<sub>2</sub> desorbing after oxidation of CO on Pt(111). *Surf. Rev. Lett.* **2**, 741–758 (1995).
13. Allers, K. H., Pfnur, H., Feulner, P. & Menzel, D. Fast reaction-products from the oxidation of CO on Pt(111) - angular and velocity distributions of the CO<sub>2</sub> product molecules. *J. Chem. Phys.* **100**, 3985–3998 (1994).
14. Gebhardt, C. R., Rakitzis, T. P., Samartzis, P. C., Ladopoulos, V. & Kitsopoulos, T. N. Slice imaging: A new approach to ion imaging and velocity mapping. *Rev. Sci. Instrum.* **72**, 3848–3853 (2001).
15. Harding, D. J., Neugeboren, J., Auerbach, D. J., Kitsopoulos, T. N. & Wodtke, A. M. Using ion imaging to measure velocity distributions in surface scattering experiments. *J. Phys. Chem. A* **119**, 12255–12262 (2015).
16. Matsushima, T. Angle-resolved measurements of product desorption and reaction dynamics on individual sites. *Surf. Sci. Rep.* **52**, 1–62 (2003).
17. Langmuir, I. Chemical reactions at low pressures. *J. Am. Chem. Soc.* **37**, 1139–1167 (1915).
18. Campbell, C. T., Ertl, G., Kuipers, H. & Segner, J. A molecular-beam investigation of the interactions of CO with a Pt(111) surface. *Surf. Sci.* **107**, 207–219 (1981).
19. Eichler, A. CO oxidation on transition metal surfaces: reaction rates from first principles. *Surf. Sci.* **498**, 314–320 (2002).
20. Acharya, C. K. & Turner, C. H. CO oxidation with Pt(111) supported on pure and boron-doped carbon: A DFT investigation. *Surf. Sci.* **602**, 3595–3602 (2008).
21. Diaz, C. et al. Chemically accurate simulation of a prototypical surface reaction: H<sub>2</sub> dissociation on Cu(111). *Science* **326**, 832–834 (2009).
22. Jørgensen, M. & Gronbeck, H. Adsorbate entropies with complete potential energy sampling in microkinetic modeling. *J. Phys. Chem. C* **121**, 7199–7207 (2017).
23. Schwarz, J. A. & Madix, R. J. Modulated beam relaxation spectrometry - its application to study of heterogeneous kinetics. *Surf. Sci.* **46**, 317–341 (1974).
24. Ross, J. A. & Jameson, D. M. Time-resolved methods in biophysics. 8. Frequency domain fluorometry: applications to intrinsic protein fluorescence. *Photochem. Photobiol. Sci.* **7**, 1301–1312 (2008).
25. Nørskov, J. K. et al. The nature of the active site in heterogeneous metal catalysis. *Chem. Soc. Rev.* **37**, 2163–2171 (2008).
26. Harding, D. J. et al. Ion and velocity map imaging for surface dynamics and kinetics. *J. Chem. Phys.* **147**, 013939 (2017).
27. Golibrzuch, K. et al. CO desorption from a catalytic surface: elucidation of the role of steps by velocity-selected residence time measurements. *J. Am. Chem. Soc.* **137**, 1465–1475 (2015).
28. Seebauer, E. & Allen, C. E. Estimating surface diffusion coefficients. *Prog. Surf. Sci.* **49**, 265–330 (1995).
29. Badan, C. et al. Step-type selective oxidation of platinum surfaces. *J. Phys. Chem. C* **120**, 22927–22935 (2016).
30. Jinnouchi, R., Kodama, K. & Morimoto, Y. DFT calculations on H, OH and O adsorbate formations on Pt(111) and Pt(332) electrodes. *J. Electroanal. Chem.* **716**, 31–44 (2014).

**Acknowledgements** A.M.W. acknowledges support from the Alexander von Humboldt Foundation. We acknowledge support from Deutsche Forschungsgemeinschaft (DFG) and the Ministerium für Wissenschaft und Kultur (MWK) Niedersachsen, and the Volkswagenstiftung under grant INST 186/952-1. C.T.C. acknowledges the Göttingen Academy of Sciences and the US National Science Foundation (grant CHE-1665077) for support.

**Reviewer information** Nature thanks E. Hasselbrink and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** J.N. and H.W.H. performed experiments with Pt(111), analysed the data and contributed to the writing of the paper. D.B. carried out experiments on Pt(332) and analysed the data. J.A. and A.K. carried out DFT calculations and performed ab initio molecular dynamics simulations. C.T.C. contributed to experimental methodology development and contributed useful arguments that led to the discovery of the reaction mechanism. D.J.A. contributed to discussions of the mechanism and to the writing of the paper. D.S. contributed to the discussion of the mechanism. D.J.H. helped build the instrument and took data. A.M.W. conceived the experiment and contributed to writing the paper. T.N.K. developed the use of slice ion imaging for reactive surface scattering, took data, developed data analysis tools and contributed to writing the manuscript.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0188-x>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to T.N.K.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## METHODS

**Experimental methods.** The instrument used in this work has been described recently for the measurements of desorption rates of CO from Pt(111) and Pd(111)<sup>26</sup>. A platinum crystal is first prepared by argon ion sputtering and annealing. The O<sub>2</sub> (normal incidence) and CO (20° from normal) pulsed molecular beams cross at the platinum surface at a selected repetition rate ratio (RRR). We used 7-bar backing pressure for the O<sub>2</sub> beam with a nozzle opening time of 70 μs. This gave an O<sub>2</sub> flux of  $1.1 \times 10^{11}$  molecules per pulse. A 5% mixture of CO in He at a total backing pressure of 7 bar and a nozzle opening time of 25 μs gave a CO flux of  $2.2 \times 10^{10}$  molecules per pulse. The diameter of both molecular beams at the surface was 3 mm. The O<sub>2</sub> beam was always fired at least 500 μs before the CO beam. The CO pulse initiates the reaction and is consumed before the next CO pulse arrives. A single CO pulse alters the oxygen coverage negligibly. CO<sub>2</sub> products were photoionized by a nine-photon non-resonant absorption process induced by the focused output (20 cm focal length lens) of a Coherent Astrella Ti:sapphire laser producing 1–5 mJ per pulse with a pulse duration of 100 fs at a wavelength of 800 nm. The ion camera system was time-gated at a controlled delay with respect to the ionizing laser pulse to restrict detection to  $m/z = 44$  (CO<sub>2</sub><sup>+</sup>). The laser ionization pulse was synchronized to the CO dosing pulse; the delay between these two pulses is used to analyse the kinetics of the reaction and at each delay, ion images are recorded using modified DaVis software (LaVision). Angular distributions are obtained by physically scanning the focal point of the ionizing laser beam back and forth in front of the reacting surface while accumulating the ion image. For kinetics measurements, the laser position was fixed and reaction products were detected close to the surface normal. The RRR value corresponds to a specific oxygen coverage, which is measured in a separate titration experiment in which the O<sub>2</sub> beam is suddenly turned off and the CO<sub>2</sub> signal is recorded until all oxygen is removed from the surface. This is calibrated by titration against an oxygen-saturated platinum surface. See Methods section ‘Total oxygen coverage, [O<sub>a</sub>], from a titration experiment’.

**Conversion of the time-dependent raw data to the kinetic trace.** In measuring kinetics, the fundamental quantity of interest is the kinetic trace, defined as the product flux as a function of the reaction time. Of course, kinetic traces for reactant loss and transient populations of intermediates can also be of interest, but for simplicity we consider here only the appearance of the product.

The oxidation of CO at platinum exemplifies the challenges in obtaining this quantity experimentally. As the reaction is initiated by CO adsorption, it is usual to use pulsed molecular beam methods. Time is required for CO oxidation to take place and for the product, CO<sub>2</sub>, to desorb from the surface. The time at which the product appears at a detector is then recorded with respect to the reaction-initiating CO pulse. Hence, the time dependence of the raw signal obtained in this experiment is composed of three contributions: the time required for the CO pulse to travel to the surface, the reaction time at the surface, and the time required for the products to travel to the ionizing spot created by the laser. Only the second time period is of interest in a kinetics experiment. The first is identical for the two reaction channels and therefore only relevant to the determination of the time at which the reaction is initiated. Finally, the third time period is different for each of the two reaction channels and must be measured.

In velocity-resolved kinetics we measure ion images such as those shown in Extended Data Fig. 1a for each time delay between the CO and ionizing laser pulses. Here, the signal intensity is proportional to the CO<sub>2</sub> number density; as such, the hyperthermal channel (red box) is only weakly seen.

Multiplying the intensity by the velocity converts the density image to a CO<sub>2</sub> product flux image (Extended Data Fig. 1b). We then select velocity integration windows (green/blue and red rectangles in Extended Data Fig. 1) and record the average intensity versus the delay between the CO pulse and the ionizing laser pulse, producing raw kinetic traces. Extended Data Fig. 1c, d shows how this appears as a density signal and as flux, respectively.

Extended Data Fig. 2a shows the velocity distribution of the product CO<sub>2</sub>, and can be decomposed into two contributions: thermal (blue) and hyperthermal (red). The velocity windows used for integration in Extended Data Fig. 1 are also shown (blue and red hatched areas). The kinetic traces are then scaled to represent the actual area under the fit to the velocity distribution (blue and red shaded areas).

A measurement of the angular distribution (Extended Data Fig. 2b, c) shows that the thermal CO<sub>2</sub> follows a cosine distribution, whereas the hyperthermal CO<sub>2</sub> follows a cos<sup>8</sup> distribution, in agreement with previous reports<sup>11–13,31</sup>.

The velocity distribution in Extended Data Fig. 2a includes only ions observed within 3° of the surface normal. In order for the kinetic traces to represent the total flux over all angles, we must scale up the kinetic trace of the thermal channel by an angular factor:

$$\frac{\int_0^{3^\circ} \cos^8 \theta \sin \theta d\theta}{\int_0^{90^\circ} \cos^8 \theta \sin \theta d\theta} \div \frac{\int_0^{3^\circ} \cos \theta \sin \theta d\theta}{\int_0^{90^\circ} \cos \theta \sin \theta d\theta} = 4.7$$

Note that in Extended Data Fig. 1d, the signal of the hyperthermal channel appears earlier than that of the thermal channel. This is a consequence of the reduced flight time of hyperthermal CO<sub>2</sub> to the ionizing laser spot, relative to that of the thermal CO<sub>2</sub>. By subtracting the flight time using the average velocity associated with the integration window of each channel (12 and 35 μs, respectively), we can correct this. See Extended Data Fig. 2d, in which we have also shifted the zero of time to the onset of the CO adsorption. The temporal shape of the CO dosing pulse is measured as described in ref. 27.

**Effective lifetimes as a function of oxygen coverage and step-dependent desorption.** We perform a preliminary and approximate data analysis at low oxygen coverages, deriving effective lifetimes for each reaction channel, which leads to insights into the reaction mechanism. Extended Data Fig. 3a shows an example of the analysis. Here, the effective lifetime of the two reaction channels is derived through a simple exponential fit to the kinetic traces, convoluted over the time profile of the pulsed molecular beam.

The lifetimes obtained from the fits are effective values,  $\tau_{\text{eff}} = 1/k_{\text{eff}}$ , which depend on the total oxygen coverage [O<sub>a</sub>]. The CO doses here are so small (less than 0.1% of a monolayer) that [O<sub>a</sub>] is independent of time during all transients. Referring to Extended Data Fig. 3b, it can be shown that  $k_{\text{eff}} = k_0 + k_2[\text{O}_a]$ , where  $k_2$  is the rate constant of the oxidation reaction and  $k_0$  is a non-reactive competitive process removing the reactant CO, for example CO desorption.

With this mechanism, a simple set of differential equations can be solved analytically. This enables us to define the effective lifetime:

$$-\frac{d[\text{CO}_a]}{dt} = k_0[\text{CO}_a] + k_2[\text{CO}_a][\text{O}_a] = [\text{CO}_a](k_0 + k_2[\text{O}_a]) = [\text{CO}_a]k_{\text{eff}}$$

$$[\text{CO}_a]_t = [\text{CO}_a]_0 e^{-k_{\text{eff}}t}$$

$$\text{Flux}(\text{CO}_2) = k_2[\text{O}_a][\text{CO}_a]_t = k_2[\text{O}_a][\text{CO}_a]_0 e^{-k_{\text{eff}}t}$$

By varying the repetition rates of the two pulsed beams, we vary the fluxes of CO and O<sub>2</sub> and thereby the coverage of oxygen. Note that the use of repetition rate gives us very precise control of the fluxes. Methods section ‘Total oxygen coverage, [O<sub>a</sub>], from a titration experiment’ shows how the oxygen coverage is derived. In this way, we can observe a linear dependence between  $k_{\text{eff}}$  and the O<sub>2</sub> flux at low O<sub>2</sub> fluxes. Extended Data Fig. 3c shows an example of how  $k_{\text{eff}}^{\text{thermal}}$  depends on the oxygen flux at several surface temperatures.

For each temperature, we extract  $k_0^{\text{thermal}}$  from the zero-coverage intercept of the linear fit. These are shown as filled circles in Extended Data Fig. 4, where they are compared to literature rate constants for CO desorption from Pt(111) steps (plus signs and filled circles) and terraces (triangles). The values of  $k_0^{\text{thermal}}$  derived here are in excellent agreement with desorption rate constants for CO adsorbed at steps.

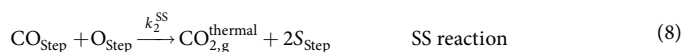
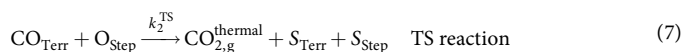
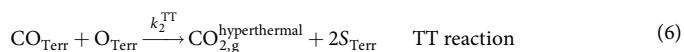
Similar analysis for the hyperthermal channel on Pt(332) also reveals a non-reactive competitive channel that removes CO,  $k_1^{\text{hyperthermal}}$ , which is consistent with CO diffusion across terraces together with desorption.

**DFT calculation of minimum energy path for the terrace reaction.** We performed DFT calculations within the generalized gradient approximation using the aimsChain tool within the FHI-aims package<sup>32</sup> to map out the minimum energy path of the CO oxidation reaction on Pt(111) terraces. The calculation used the RPBE functional in the spin-unpolarized (restricted) formalism with van der Waals corrections<sup>33</sup>. Other details of the calculations are as follows: A p(3 × 3) cell with 4 layers with an optimized lattice constant of 4.00 Å was used as a model for the Pt(111) surface. To find the initial and final state for the minimum energy path (MEP) calculation, we carried out geometry optimizations in which just CO and oxygen atoms were allowed to move and all surface atoms were restricted to their equilibrium positions. The optimizations were stopped when the force on every atom was <0.01 eV Å<sup>-1</sup>. The minimum energy configurations found in this way were also checked against tight basis sets. For both MEP and geometry optimizations, the reciprocal space was sampled with a 4 × 4 × 1 *k*-point grid, and to account for the electron occupation a Gaussian function with a width of 0.2 eV was used. The MEP calculations were performed using the string method<sup>34</sup> with light basis sets. To sample the MEP we have taken 15 images. The convergence of the pathway was reached when the forces between every image were smaller than 0.15 eV Å<sup>-1</sup>. To determine the transition state more accurately, we used a force criterion of 0.05 eV Å<sup>-1</sup> for the climbing image.

Extended Data Fig. 5 shows that the reaction is highly exoergic and passes over a substantial reaction barrier. The structure of the transition state is shown in the inset and is still far from the structure of the product, similar to that found in previous calculations<sup>19,20</sup>. This so-called ‘early barrier’ indicates that a substantial fraction of the reaction energy will be channelled into vibrational excitation of the product, CO<sub>2</sub>.

We performed ab initio molecular dynamics calculations using the system and functional defined above, starting at the transition state determined from the MEP calculations to determine the translational energy of the product CO<sub>2</sub>. The results from about 10 trajectories show that the values of the final translational energy are between 660 meV and 715 meV for the flexible slab at  $T = 0$  K, and between 720 meV and 920 meV for the rigid slab. The results of this electronically adiabatic calculation are consistent with those of experiments, in which molecular energy may also be shared with the electron–hole pairs of the metal.

**Kinetic modelling of the velocity-resolved data for CO oxidation on Pt(111).** *Reaction scheme.* The kinetic model is based on the following reaction scheme for CO oxidation:



where  $S_{\text{Terr,Step}}$  are the free sites expressed in monolayers. Because the step density is less than 1% on our Pt(111) sample, we omit reaction (1) and consider only CO adsorption at terraces. We include reaction (1) for completeness and it will be important for our treatment of Pt(332). The diffusion from steps back to terraces has been omitted; as we see different temporal behaviour of the two kinetic traces in Extended Data Fig. 1, we conclude that the two CO species are not in fast equilibrium with each other. The diffusion from steps back onto the terrace must therefore be slow compared to the other processes.

The rate equations describing this mechanism are as follows:

$$\begin{aligned} \frac{d[\text{O}_{\text{Step}}]}{dt} &= -(k_1^{\text{TS}}[\text{CO}_{\text{Terr}}] + k_2^{\text{SS}}[\text{CO}_{\text{Step}}]) \frac{[\text{O}_{\text{Step}}]}{d_{\text{Step}}} \\ \frac{d[\text{CO}_{\text{Step}}]}{dt} &= - \left( k_0^{\text{S}} + k_2^{\text{SS}} \frac{[\text{O}_{\text{Step}}]}{d_{\text{Step}}} \right) [\text{CO}_{\text{Step}}] + k_1^{\text{TS}} [\text{CO}_{\text{Terr}}] \frac{d_{\text{Step}} - [\text{O}_{\text{Step}}] - [\text{CO}_{\text{Step}}]}{d_{\text{Step}}} \\ \frac{d[\text{CO}_{\text{Terr}}]}{dt} &= - \left( k_0^{\text{T}} + k_1^{\text{TS}} \frac{d_{\text{Step}} - [\text{O}_{\text{Step}}] - [\text{CO}_{\text{Step}}]}{d_{\text{Step}}} + k_2^{\text{TS}} \frac{[\text{O}_{\text{Step}}]}{d_{\text{Step}}} \right) [\text{CO}_{\text{Terr}}] + k_2^{\text{TT}} \frac{[\text{O}_{\text{Terr}}]}{d_{\text{Terr}}} \\ |\text{Flux}(\text{CO}_{2,\text{g}}^{\text{thermal}})| &= k_2^{\text{TS}} [\text{CO}_{\text{Terr}}] \frac{[\text{O}_{\text{Step}}]}{d_{\text{Step}}} + k_2^{\text{SS}} [\text{CO}_{\text{Step}}] \frac{[\text{O}_{\text{Step}}]}{d_{\text{Step}}} \\ |\text{Flux}(\text{CO}_{2,\text{g}}^{\text{hyperthermal}})| &= k_2^{\text{TT}} [\text{CO}_{\text{Terr}}] \frac{[\text{O}_{\text{Terr}}]}{d_{\text{Terr}}} \end{aligned}$$

where the square brackets refer to concentration in monolayers of surface-adsorbed reactants and products with respect to the entire surface,  $d_{\text{Step}}$  and  $d_{\text{Terr}}$  are the concentrations of the step and terrace sites expressed in monolayers with respect

to the entire surface that is,  $d_{\text{Step}} + d_{\text{Terr}} = 1$  monolayer,  $|\text{Flux}(\text{CO}_{2,\text{g}})|$  refers to the absolute value of the flux of CO<sub>2,g</sub> gas phase product molecules. All rate constants  $k$  are in units of  $\text{s}^{-1}$ .

We stress here that when probing the gas-phase product from a surface reaction, the flux of this species must be measured and not the rate of change of the product density,  $d[\text{CO}_{2,\text{g}}]/dt$ . A simple dimensional analysis of the rate law clarifies this point:

$$k_2^{\text{TT}} [\text{CO}_{\text{Terr}}] \frac{[\text{O}_{\text{Terr}}]}{d_{\text{Terr}}} \quad \text{has units of} \quad \frac{\text{monolayers}}{\text{s}} \propto \frac{\text{particles}}{\text{cm}^2 \text{ s}}$$

while

$$\frac{d[\text{CO}_{2,\text{g}}]}{dt} \quad \text{has units of} \quad \frac{\text{particles}}{\text{cm}^3 \text{ s}}$$

Flux has, by definition, units of particles  $\text{cm}^{-2} \text{ s}^{-1}$ .

The physical picture that further confirms this is the fact that as the particles desorb from a two-dimensional surface area, after some time they occupy a certain volume. This volume is proportional to the velocity. Measuring simply the density of the particles is insufficient, as high-speed and low-speed particles will have ‘expanded’ into different volumes.

Ion imaging can therefore function as a flux detector and, as such, is ideally suited for the measurement of kinetics at surfaces that result in gas-phase products. *Numerical solution to the kinetic model.* We derive a system of ordinary differential equations for the reaction scheme above and solve them numerically. The initial values for the concentration of oxygen, that is,  $[\text{O}_{\text{Terr}}]$  and  $[\text{O}_{\text{Step}}]$ , are provided as an initial guess from the titration data (see Methods section ‘Validation of  $[\text{O}_{\text{Terr}}]$  and  $[\text{O}_{\text{Step}}]$  from the numerical solution’). The concentration of free sites,  $[S_{\text{Terr}}]$  and  $[S_{\text{Step}}]$ , are expressed in units of monolayers, that is,  $[S_{\text{Terr}}] = 0.9975$  monolayers and  $[S_{\text{Step}}] = 0.0025$  monolayers for an empty surface and a step density of 0.25% as determined from AFM measurements. The rate constants for desorption from steps and terraces are known from previous experiments, and the rate constant for diffusion is estimated from literature values. An initial guess for the rate constants of the reaction is provided from an analysis of  $k_{\text{eff}}$  (see Methods section ‘Effective lifetimes as a function of oxygen coverage and step-dependent desorption’). The adsorption of CO from the incoming beam is then introduced as a perturbation to this system.

*Optimization process.* The flux of CO<sub>2</sub> into the thermal and hyperthermal channels as a function of time was then obtained by solving the ordinary-differential-equation system using LSODA from the FORTRAN77 library ODEPACK<sup>35</sup>. This flux was compared to the experimental data and the residuals were minimized in a recursive optimization. We used four global parameters,  $k_1^{\text{TS}}$ ,  $k_2^{\text{TT}}$ ,  $k_2^{\text{TS}}$  and  $k_2^{\text{SS}}$ , and two local parameters per dataset, the concentration of  $[\text{O}_{\text{Terr}}]$  and  $[\text{O}_{\text{Step}}]$ , to perform optimization with the Nelder–Mead method. Following ref.<sup>36</sup>,  $[\text{O}_{\text{Step}}]$  can occupy every second (fcc) platinum-step site, therefore the maximum of  $[\text{O}_{\text{Step}}]$  was restricted to  $< 0.5d_{\text{g}}$ .

Using these four global parameters, the kinetic model was able to properly reproduce 18 sets of thermal and hyperthermal flux at 7 temperatures, a total of 252 kinetic traces. Examples are shown in Extended Data Fig. 6.

*Total oxygen coverage,  $[\text{O}_a]$ , from a titration experiment.* The total amount of oxygen on the surface,  $[\text{O}_a]$ , could be determined in a titration experiment. First, a surface fully covered with oxygen (0.25 monolayers) was continuously dosed with CO and the total CO<sub>2</sub> yield was measured. Next, the CO and O<sub>2</sub> beams were run at a value of RRR used when measuring kinetic traces; this established a steady-state oxygen coverage. After several minutes, the O<sub>2</sub> beam was suddenly turned off and the total CO<sub>2</sub> yield was again measured. This CO<sub>2</sub> yield was then compared to the yield obtained from a fully covered surface, allowing us to relate RRR to the resulting steady-state oxygen coverage. The result is shown as black squares in Extended Data Fig. 7a.

*Validation of  $[\text{O}_{\text{Terr}}]$  and  $[\text{O}_{\text{Step}}]$  from the numerical solution.* To test whether the values for  $[\text{O}_{\text{Terr}}]$  and  $[\text{O}_{\text{Step}}]$  obtained from the numerical model are reasonable, we compare their sum to the total amount of oxygen on the surface as obtained from the titration experiments just described. The kinetic model result for  $[\text{O}_a]$  is plotted as red dots in Extended Data Fig. 7a. The good agreement between the prediction of the numerical solution and the titration determined values of oxygen coverage validates the way in which the numerical model deals with the oxygen coverage.

In addition to this, we want to confirm that the partitioning of  $[\text{O}_a]$  into  $[\text{O}_{\text{Terr}}]$  and  $[\text{O}_{\text{Step}}]$  is reasonable. The equilibrium between the two oxygen species is calculated using the canonical partition function for oxygen atoms distributed among step and terrace sites using a given energy difference as described in Methods section ‘Equilibrium calculation of oxygen-atom populations at steps and terraces’. The step coverage is defined as

$$\theta_{\text{Step}} = \frac{[\text{O}_{\text{Step}}]}{0.5d_{\text{Step}}}$$

with the step density  $d_{\text{step}} = 0.0025$ . As explained in Methods section ‘Optimization process’, one oxygen atom occupies every second binding site, therefore the factor of 0.5 is used.

A plot comparing the numerical results to the equilibrium expectation is shown in Extended Data Fig. 7b. Here the relative binding energy of oxygen at steps and terraces was varied between 0.05 and 0.16 eV. Theoretical calculations using DFT suggest that step oxygen is more strongly bound than terrace oxygen, by about 0.26 eV<sup>30,37</sup>.

Therefore, the data suggests that the equilibrium distribution of oxygen atoms between steps and terraces is not established. Given the slow diffusion speeds on Pt(111) at our temperatures and the large average terrace size of about 400 atoms, this is conceivable. The distribution on Pt(332), which has an average terrace length of only five atoms, appears closer to or at equilibrium (see Methods section ‘Kinetic studies of CO oxidation on Pt(332)’).

**Equilibrium calculation of oxygen-atom populations at steps and terraces.** In developing the numerical model for the oxidation of CO on platinum, we must account for the partitioning of oxygen atoms between steps and terraces. Theoretical calculations using DFT suggest that step-bound oxygen is more strongly bound than terrace-bound oxygen (by about 0.26 eV<sup>30,37</sup>). One limiting approach to this problem is equilibrium partitioning.

Whereas step-bound oxygen is favoured by energetic considerations, terrace-bound oxygen atoms are favoured by entropy. To describe this, we have used a simple expression for the equilibrium constant which can be derived from the canonical partition function of non-interacting oxygen atoms bound at two sites (steps and terraces):

$$Q(n_{\text{step}}, n_{\text{terr}}, n_{\text{O}}, E_{\text{TS}}, \beta) = \sum_{m=0}^{n_{\text{step}}} \binom{n_{\text{step}}}{m} \binom{n_{\text{terr}}}{n_{\text{O}}-m} e^{-\beta((n_{\text{O}}-m)E_{\text{TS}}-E_0)}$$

Here, binomial coefficients are used to calculate the number of ways to distribute  $n_{\text{O}}$  oxygen atoms over  $n$  step sites ( $n_{\text{step}}$ ) and  $n$  terrace sites ( $n_{\text{terr}}$ ). We take into account that saturated coverage of steps occurs at a 2:1 ratio of platinum to oxygen atoms, whereas saturation of terraces occurs at a 4:1 ratio.  $E_{\text{TS}}$  is the energetic stability difference between step-bound oxygen atoms and terrace-bound oxygen atoms. If we use this equilibrium partitioning to characterize our experimentally derived results by allowing  $E_{\text{TS}}$  to vary, we find a value that depends on step density. For Pt(111) we find  $E_{\text{TS}} = 0.05$  eV (see Methods section ‘Validation of  $[O_{\text{terr}}]$  and  $[O_{\text{step}}]$  from the numerical solution’), whereas for Pt(332) we find  $E_{\text{TS}} = 0.12$  eV (see Methods section ‘Kinetic studies of CO oxidation on Pt(332)’). We observe that the  $E_{\text{TS}}$  for Pt(332) is closer to that obtained from the DFT calculations<sup>30,37</sup>, which could suggest that the system on the (332) surface is closer to equilibrium.

**Kinetic studies of CO oxidation on Pt(332).** The kinetic mechanism for CO oxidation on Pt(332) is identical to that described in Methods section ‘Reaction scheme’ for Pt(111); however, here the step density (0.167) is large. Therefore, we include both reactions (1) and (2) weighted by 0.167 and 0.833, respectively. Based on our success with the treatment of the Pt(111) data, we implemented the numerical model as follows.

First, we used titration measurements as described in Methods section ‘Total oxygen coverage,  $[O_a]$ , from a titration experiment’ as a starting point for the determination of the total oxygen coverage as a function of the ratio of  $O_2$  to CO fluxes (Extended Data Fig. 7c). We then constructed limits (red shaded area in Extended Data Fig. 7c) within which the oxygen coverage could be varied in the numerical optimization to reflect the uncertainty of the titration measurements. For the initial conditions of the numerical optimization, we assume the partitioning of oxygen between steps and terraces is at equilibrium as discussed in Methods section ‘Validation of  $[O_{\text{terr}}]$  and  $[O_{\text{step}}]$  from the numerical solution’. After the fitting of the data is complete, the partitioning of the oxygen atoms at steps and terraces has also been optimized and is shown as black triangles in Extended Data Fig. 7d. These are compared to equilibrium calculations assuming different values of the binding energy difference between oxygen atoms at steps and terraces. The results obtained from the numerical fit are similar to equilibrium values when the energetic preference for step-bound oxygen atoms was 0.10–0.14 eV. Compared to the results derived from the Pt(111) experiments, this value is much closer to the theoretically estimated difference in binding energy. This is probably due to much smaller terrace sizes leading to faster equilibration compared to Pt(111).

Extended Data Fig. 8 shows examples of the quality of the fit obtained using the rate constants from Table 1. We emphasize that both reactions (7) and (8)—the terrace-step and step-step reactions, respectively—are necessary to account for the biexponential fall-off in the thermal channel. Furthermore, the magnitudes of the rate constants obtained from the Pt(111) and Pt(332) surfaces are in excellent agreement; see Fig. 3 and Table 1.

We point out that the derived activation energies for the three reactions make intuitive sense. For early-barrier reactions, the relative energies of the transition states follow closely the relative energies of the reactants. For the TT reaction, our

experimental activation energy is  $0.6 \pm 0.1$  eV. In considering the SS reaction, the step-bound oxygen atom is stabilized by about 0.15 eV (see Methods section ‘Equilibrium calculation of oxygen-atom populations at steps and terraces’) and the step-bound CO is stabilized by 0.16–0.35 eV<sup>38–40</sup>. Here there are a variety of experimental studies that give a range of results. For early-barrier reactions, the transition state is stabilized by a similar amount of energy as the reactants are stabilized. Therefore it is not surprising that the SS reaction has a similar activation energy ( $0.65 \pm 0.1$  eV) to that of the TT reaction ( $0.6 \pm 0.1$  eV). We speculate that the TS reaction and the SS reaction access the same transition state via two different reactant geometries. If true, the activation energy for the TS reaction would be lowered by the step-to-terrace relative binding energy of CO, which is 0.16–0.3 eV. This simple analysis would predict a lower activation energy for the TS reaction of anywhere between 0.49 and 0.25 eV, which is consistent with our experimentally derived value of  $0.4 \pm 0.1$  eV.

**Comparison of velocity-resolved kinetics to previous reported velocity-integrated kinetic results.** Previous surface-kinetics experiments were not able to resolve product velocities. Furthermore, only a single oxidation reaction was assumed, enabling a pseudo-first order analysis as a function of the surface temperature  $T_s$  and  $[O_a]$ , from which activation energies were derived. Extended Data Fig. 9b shows reported activation energies versus  $[O_a]$  for the reaction thought to result from CO oxidation at terraces (hollow circles and plus signs) from refs<sup>7,41</sup>.

We can use our velocity-resolved data to simulate these previous results. We integrate the kinetic trace over the product velocities and fit it as if pseudo-first-order kinetics applied (see Extended Data Fig. 9a). We do this for data at many temperatures and values of  $[O_a]$ . The apparent activation energies obtained in this way are shown as filled circles in Extended Data Fig. 9b, and there is marked agreement with the results of previous work.

We emphasize that the true oxidation kinetics involves three reactions, the activation energies of which are independent of  $[O_a]$ . This exemplifies how velocity-resolved kinetics removes a layer of averaging that confounds accurate kinetic analysis.

The interpretations of previous work suggested that the activation energy of a surface reaction could vary continuously with coverage, leading to a change of more than 0.5 eV owing to oxygen-atom occupation of neighbouring reaction sites at a distance of around 8 Å or more. This is clearly not the case. It is notable that the assumption of the single-reaction mechanism leads to this erroneous conclusion.

**Temperature dependence of the site-specific mechanism.** We used the numerical model with rate parameters that have been optimized to reproduce the velocity-resolved kinetics data to explore the behaviour of CO oxidation as a function of temperature and oxygen coverage. Extended Data Fig. 10a shows the CO oxidation efficiency (probability that an adsorbed CO molecule is converted to  $CO_2$ ). Extended Data Fig. 10b shows the relative fraction of the hyperthermal channel (TT reaction) with respect to the thermal channels (TS and SS).

We note that the dominant role of step reactions between 500 and 700 K may appear inconsistent with Fig. 3, which shows  $k_2^{\text{TT}} > k_2^{\text{SS}}, k_2^{\text{TS}}$  in this temperature range. This points out how subtle the kinetics of CO oxidation can be. Although the TT reaction has a higher rate constant, it must compete with CO diffusion to steps, which is faster than the terrace reaction at these temperatures. Furthermore, CO is preferentially adsorbed at steps; as such, it does not so easily diffuse back to the terraces once it has arrived at a step. Although it is true that step reactions have lower rate constants than do terrace reactions, CO molecules that have arrived at steps may only either react or desorb. Because desorption is rather slow, the reaction dominates. Another factor that is especially important when the total oxygen coverage is low is that oxygen atoms are preferentially found at step sites, which obviously favours reaction at steps.

Therefore, the relative importance of step and terrace reactions is strongly influenced by diffusion and desorption. Under conditions in which the TT reaction dominates, the reaction probability of conversion from CO to  $CO_2$  is small. We explain this by observing that the TT reaction can only dominate when CO desorption is faster than CO diffusion. Under such conditions, desorption is also faster than the TT reaction. We note that this is consistent with previous observations, in which only the hyperthermal channel was seen<sup>8</sup>. We have carried out similar calculations for the Pt(332) catalysts with step densities of 16.7%, and the results are nearly identical.

It is also interesting to note that many industrial catalysts are used in a temperature region in which the CO– $CO_2$  conversion efficiency is low.

**Sensitivity analysis.** The three lower panels in Extended Data Fig. 10 show how the fit to the data (represented by the total fit residual  $|\chi|$ ) depends on the three activation energies derived from our numerical model. Each two-dimensional contour plot shows sensitivity and the potential correlation of fitting error between two fitting parameters.

The total fit residual was obtained by summing over the array containing the residual at each data point for both channels under 152 different reaction



conditions (both temperature and oxygen coverage vary within the dataset). The Nelder–Mead optimization method used the same residual array as input. For each point within each contour plot, the activation energies were varied as indicated by the respective axis. The reported energy was varied by  $\pm 0.2$  eV in steps of 0.013 eV. The pre-exponential factors were adjusted to compensate for the effect of the change in activation energy in the middle of the temperature range of the experiments (320 °C). In other words, when varying the activation energy, the rate constant at 320 °C was kept constant by compensating the prefactor.

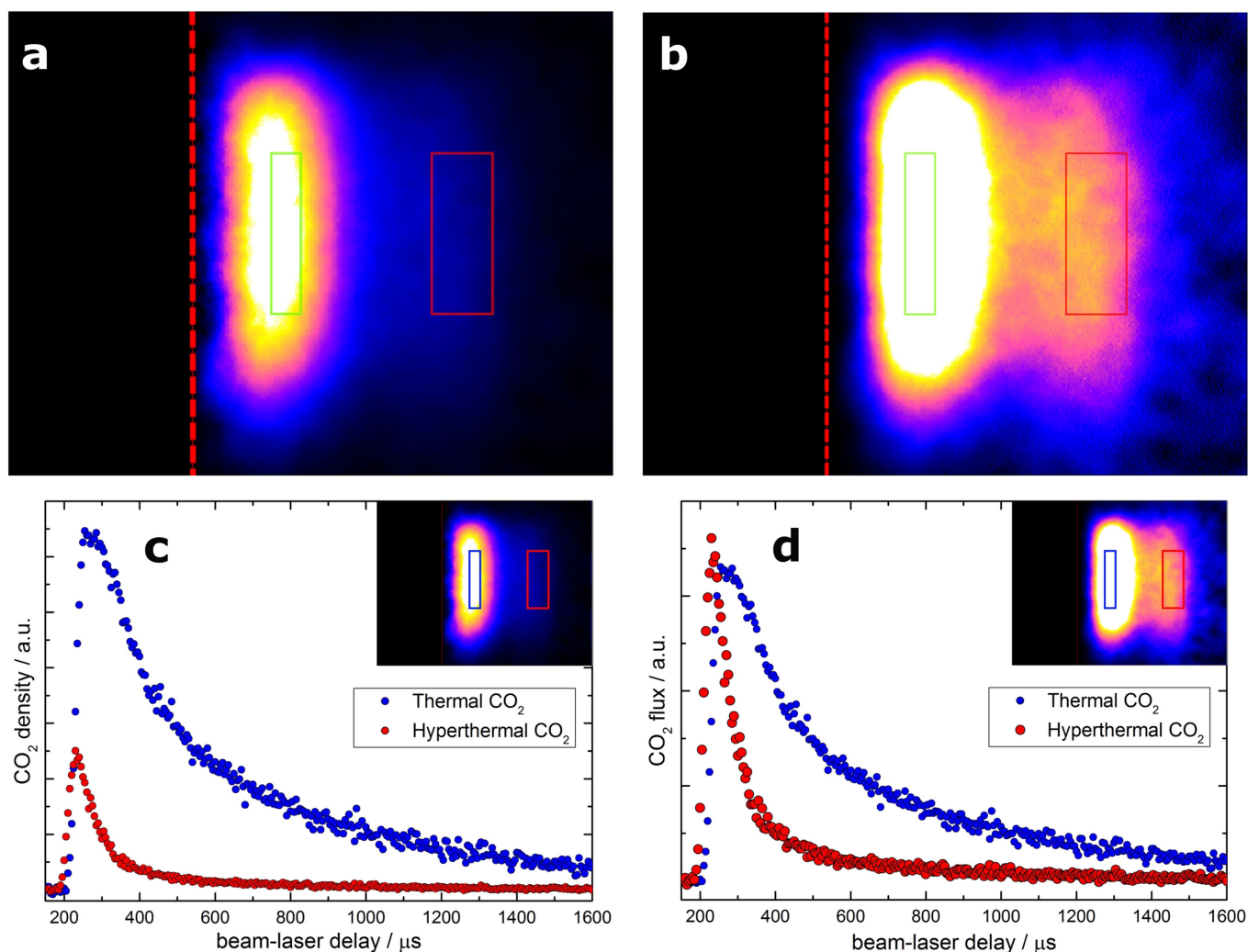
We can make a number of observations from these three plots. First, the residual is more sensitive to changes of  $E_a^{\text{TS}}$  than it is to the other two activation energies. This can be seen from the strong gradient along the horizontal direction in Extended Data Fig. 10c and the large curvature along the vertical axis in Extended Data Fig. 10d. Second,  $E_a^{\text{TT}}$  (the vertical axis of Extended Data Fig. 10c) and  $E_a^{\text{SS}}$  (the horizontal axis of Extended Data Fig. 10d) are ineffective at compensating fit error when  $E_a^{\text{TS}}$  is shifted away from its optimal value. Hence both plots appear as long and narrow valleys aligned with one of the two axes. This reflects the small correlation between these fitting parameters; the slant in Extended Data Fig. 10d represents a weak correlation in fitting error between  $E_a^{\text{TS}}$  and  $E_a^{\text{SS}}$ . Extended Data Fig. 10e shows the sensitivity of  $E_a^{\text{TT}}$  and  $E_a^{\text{SS}}$  and correlation error involved in the fitting. From this plot we can see that the reported activation energies indeed reflect the minimum residual and that the residuals depend approximately equally on  $E_a^{\text{TT}}$  and  $E_a^{\text{SS}}$ . The approximately circular form of this contour plot indicates that the two fitting degrees of freedom are approximately uncorrelated with one another and that an optimized fit is sensitive to both activation energies.

On the basis of this analysis, we suggest an uncertainty of 0.05–0.1 eV for  $E_a^{\text{TS}}$  and an uncertainty of 0.1–0.2 eV for  $E_a^{\text{TT}}$  and  $E_a^{\text{SS}}$ .

**Data availability.** The data that support the findings of this study are available from the corresponding author upon reasonable request.

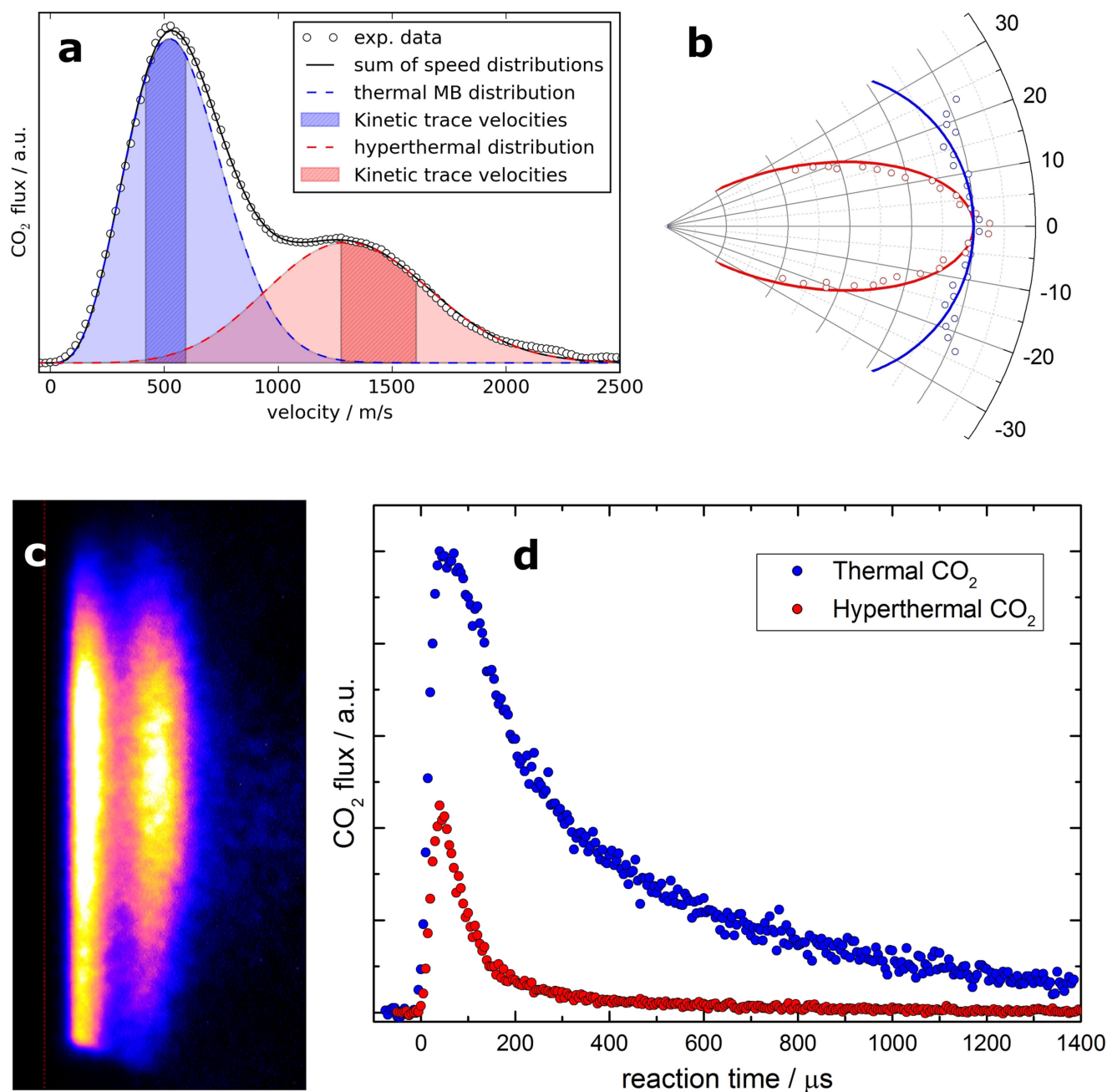
31. Poehlmann, E., Schmitt, M., Hoinkes, H. & Wilsch, H. Velocity distributions of carbon dioxide molecules from the oxidation of CO on Pt(111). *Surf. Sci.* **287–288**, 269–272 (1993).
32. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
33. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).
34. E, W., Ren, W. & Vanden-Eijnden, E. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.* **126**, 164103 (2007).
35. Hindmarsh, A. C. in *Scientific Computing* (eds Stepleman, R. S. et al.) 55–64 (North-Holland, Amsterdam, 1983).
36. Gland, J. L. & Korchak, V. N. The adsorption of oxygen on a stepped platinum single crystal surface. *Surf. Sci.* **75**, 733–750 (1978).
37. Nagoya, A., Jinnouchi, R., Kodama, K. & Morimoto, Y. DFT calculations on H, OH and O adsorbate formations on Pt(322) electrode. *J. Electroanal. Chem.* **757**, 116–127 (2015).
38. Borodin, D. *Probing Reactions at Surfaces using Ion Imaging—CO Oxidation at Atomically Flat and Stepped Surfaces of Platinum and Palladium*. MSc thesis, Georg-August Univ. Göttingen (2017).
39. Reutt-Robey, J. E., Doren, D. J., Chabal, Y. J. & Christman, S. B. Microscopic CO diffusion on a Pt(111) surface by time-resolved infrared spectroscopy. *Phys. Rev. Lett.* **61**, 2778–2781 (1988).
40. Lin, T. H. & Somorjai, G. A. Modulated molecular beam scattering of CO and NO from Pt(111) and the stepped Pt(557) crystal surfaces. *Surf. Sci.* **107**, 573–585 (1981).
41. Gland, J. L. & Kollin, E. B. Carbon monoxide oxidation on the Pt(111) surface: Temperature programmed reaction of coadsorbed atomic oxygen and carbon monoxide. *J. Chem. Phys.* **78**, 963 (1983).
42. Janda, K. C. et al. Direct measurement of velocity distributions in argon beam–tungsten surface scattering. *J. Chem. Phys.* **72**, 2403 (1980).
43. Verheij, L. K., Lux, J., Anton, A. B., Poelsema, B. & Comsa, G. A molecular beam study of the interaction of CO molecules with a Pt(111) surface using pulse shape analysis. *Surf. Sci.* **182**, 390–410 (1987).





**Extended Data Fig. 1 | Kinetics of CO<sub>2</sub> formation.** **a, b,** Ion images of the product CO<sub>2</sub>. **a,** Raw data after background subtraction, with the intensity of each pixel proportional to the density of CO<sub>2</sub>. The pixel distance from the laser position (dashed red line) is proportional to the velocity, that is, velocity increases to the right. **b,** Data after multiplying each pixel by its corresponding velocity, with the resulting pixel intensity proportional to the flux and the hyperthermal channel then becoming more readily apparent. The red and green rectangles indicate typical velocity integration windows used to produce kinetic traces for the thermal and hyperthermal

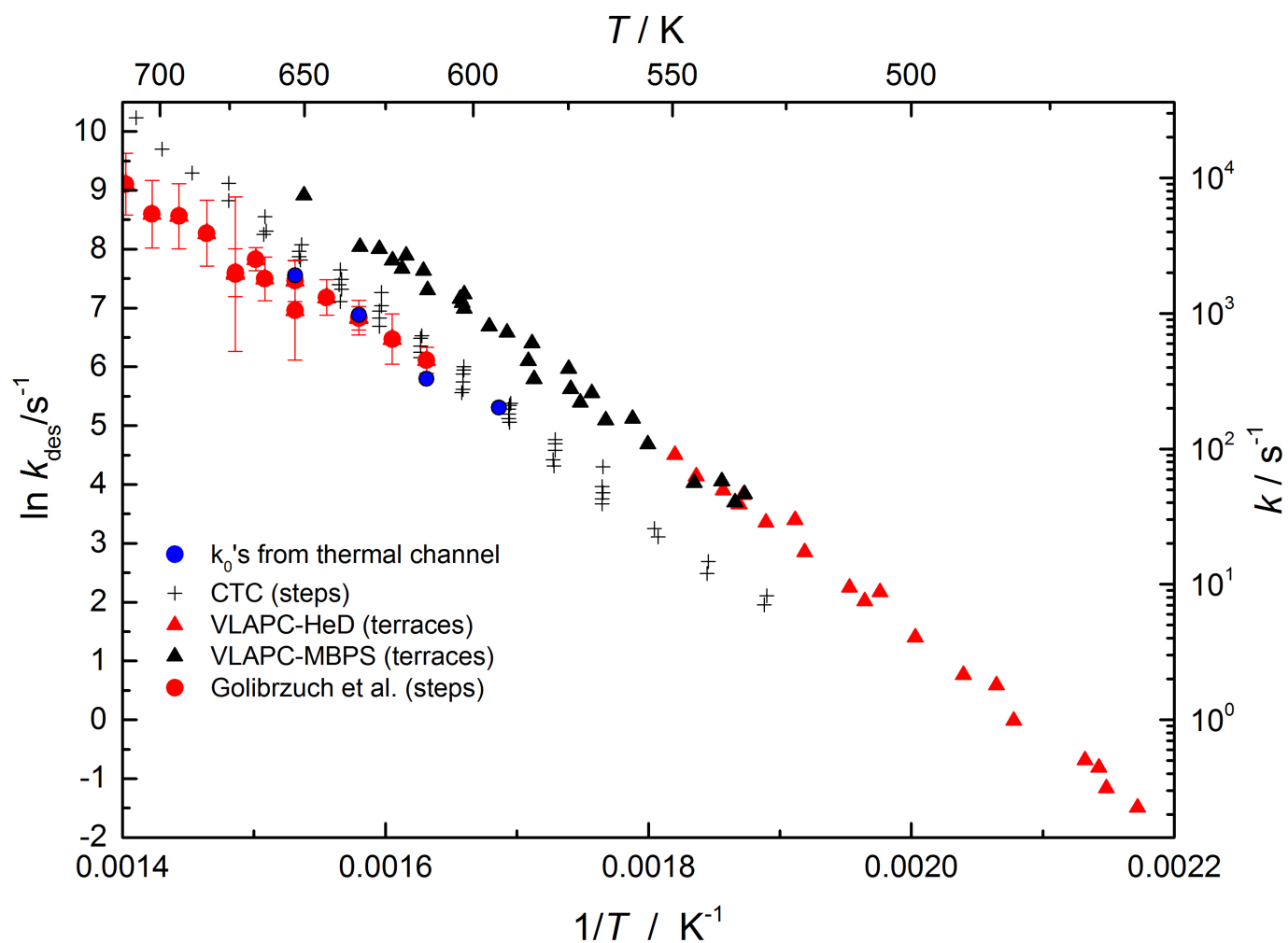
channels. The vertical span of the boxes subtends  $\pm 3^\circ$  around the surface normal. This small angular range justifies using rectangular integration windows. For defining kinetic traces, we take the average intensity within the rectangles as the product flux normal to the surface. **c, d,** Raw kinetic trace data. **c,** Average density of the CO<sub>2</sub> product integrated over the blue (thermal channel) and red (hyperthermal channel) rectangular areas of the ion image (inset). **d,** Same data converted to flux. This kinetic trace and the ion images above were measured at 350 °C, a time-averaged CO flux of  $2.2 \times 10^{12} \text{ s}^{-1} \text{ cm}^{-2}$  and a time-averaged O<sub>2</sub> flux of  $1.1 \times 10^{13} \text{ s}^{-1} \text{ cm}^{-2}$ .



**Extended Data Fig. 2 | Branching between thermal and hyperthermal reaction channels.** **a**, Product speed distribution. The branching ratio between the thermal and hyperthermal channels is obtained from this distribution, calculated from the ion image in Extended Data Fig. 1. The thermal product channel is fit to a Maxwell-Boltzmann function ( $T_{\text{trans}} = 483$  K; dashed blue line) and the hyperthermal channel is fit to a flowing Maxwell-Boltzmann function<sup>42</sup> ( $T_{\text{trans}} = 894$  K,  $\alpha = 190$  meV; dashed red line). The velocities used for extraction of the kinetic traces

are indicated by the hatched areas. **b**, The angular distribution for the ion image shown in **c**. **c**, The flux-corrected ion image that shows an angular distribution subtending  $\pm 25^\circ$  around the surface normal. The angular distribution was measured at a surface temperature of  $400^\circ\text{C}$ . **d**, The kinetic trace. We obtain the product flux as a function of reaction time for two channels with different speed and angular distributions (details are given in the text). It is now clear that the hyperthermal channel is much weaker than the thermal channel.

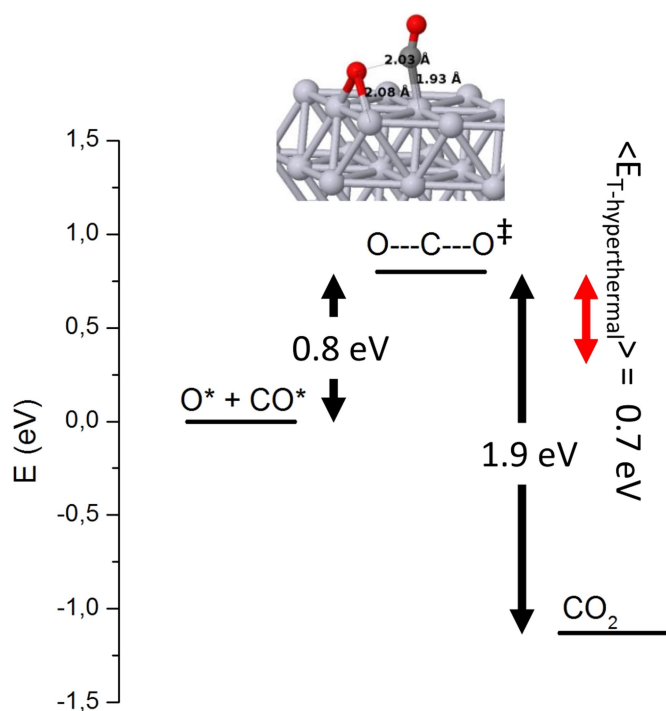




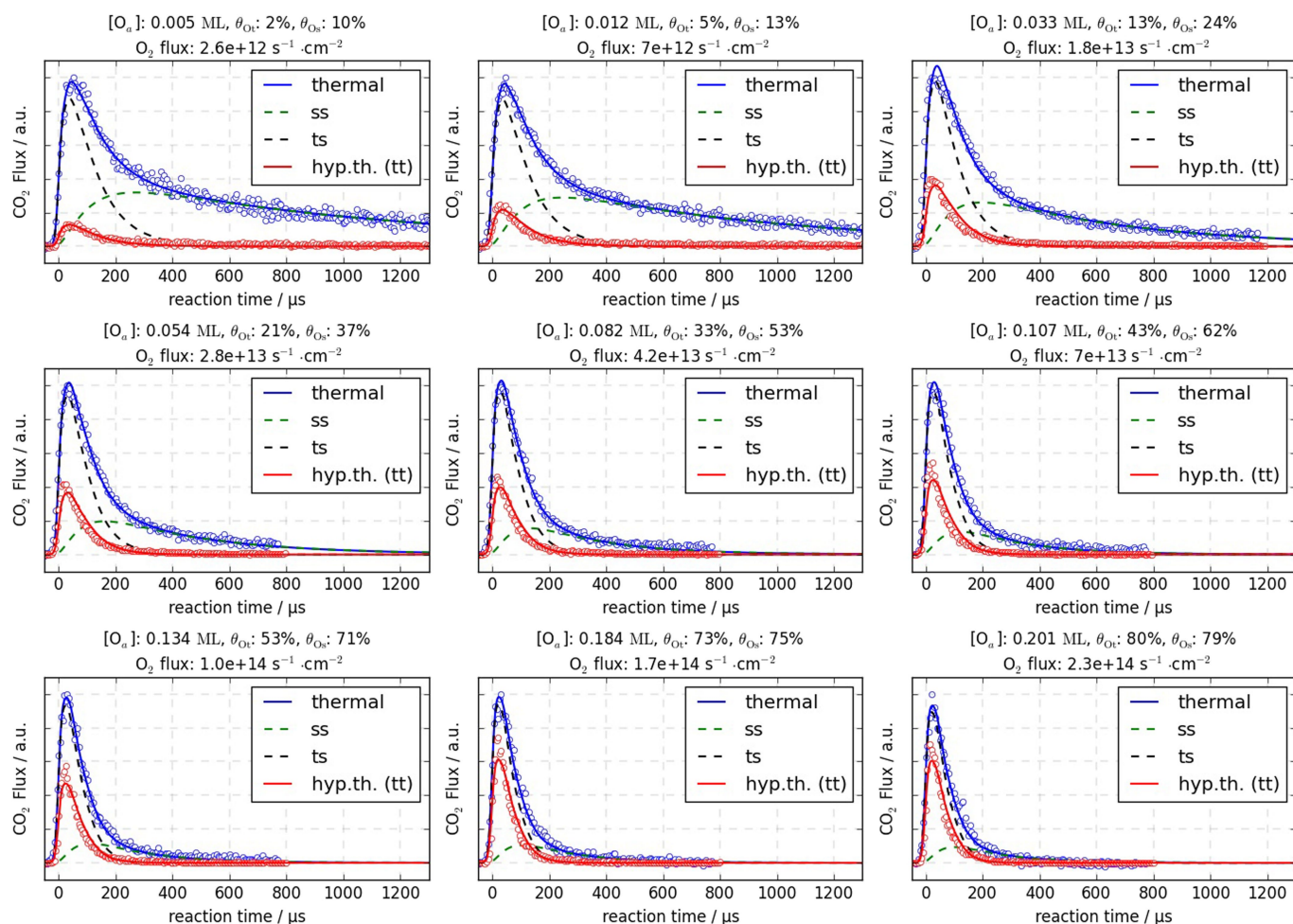
**Extended Data Fig. 4 | Rate constants for CO desorption from Pt(111).** Rate constants for desorption of CO from Pt(111) terraces (black and red triangles) are taken from ref. <sup>43</sup>. The plus signs are rate constants for desorption of CO from steps; see ref. <sup>18</sup>. The filled red circles show rate

constants for desorption of CO from steps as measured previously<sup>27</sup>. The zero-coverage rate constants from Extended Data Fig. 3 are shown as filled blue circles.



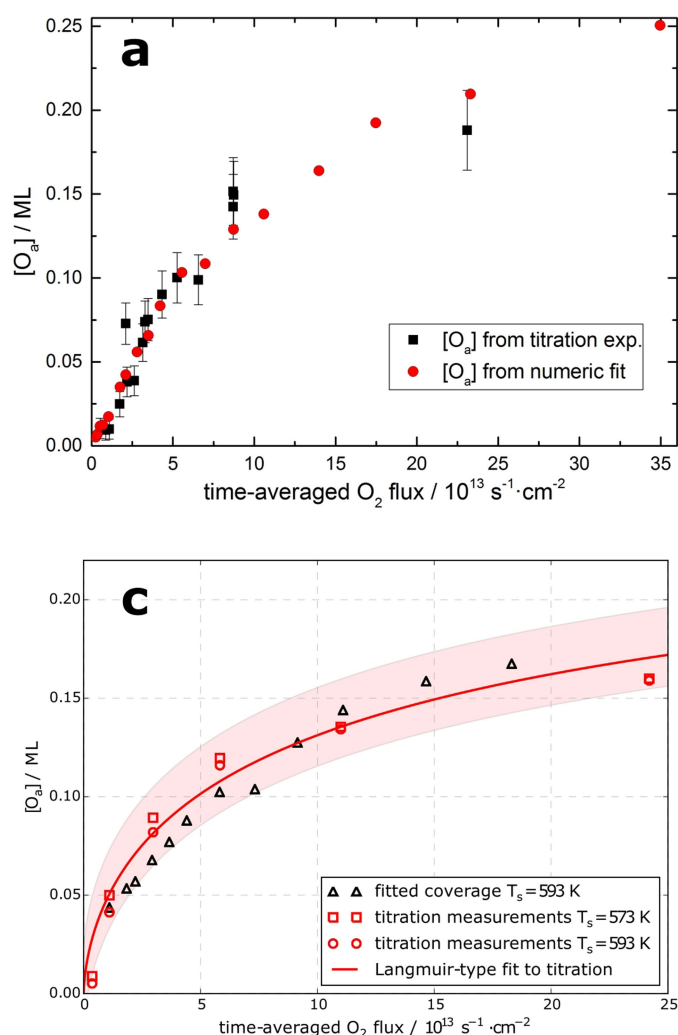


**Extended Data Fig. 5 | Points along the reaction coordinate of the terrace-terrace reaction.** The transition state resembles reactants; it is a so-called 'early-barrier' reaction. Such reactions channel the energy released in the reaction primarily into product vibration. Experimentally we observe that approximately 20% of the barrier-height energy (0.38 eV) appears as product transitional energy in the hyperthermal channel. We take this as evidence that the hyperthermal reaction takes place on platinum terraces.



**Extended Data Fig. 6 | Examples of the kinetic model fit to experimentally derived kinetic traces for Pt(111).** Nine out of 18 (every second) kinetic traces at 340 °C are shown. Similar fit quality was obtained for 18 plots for seven temperatures between 290 and 350 °C. Information

on the total oxygen coverage  $[O_a]$ , the fractional coverage on steps and terraces ( $\theta_{Ot}$  and  $\theta_{Os}$ , respectively) and the time-averaged  $O_2$  flux is shown above each plot. The time-averaged CO flux was  $2.2 \times 10^{12} \text{ s}^{-1} \text{ cm}^{-2}$ .



### Extended Data Fig. 7 | Oxygen coverage on platinum surfaces.

**a**, A comparison of the values of total adsorbed oxygen  $[O_a]$  obtained from titrations on Pt(111) with  $[O_a]$  values obtained from the numerical solution. The black squares with estimated uncertainty (1 s.d.) are the total amount of oxygen on the surface,  $[O_a]$ , obtained from titration measurements. The red dots show  $[O_a] = [O_{\text{Terr}}] + [O_{\text{Step}}]$  obtained from the numerical solution at  $340^\circ\text{C}$ . The time-averaged  $\text{CO}$  flux was  $2.2 \times 10^{12} \text{ s}^{-1} \text{ cm}^{-2}$ . **b**, The fractional step coverage ( $\theta_{os}$ ) on Pt(111) from the numerical solution compared to the partition function simulation. The black squares are the result of the numerical fit shown in Extended Data

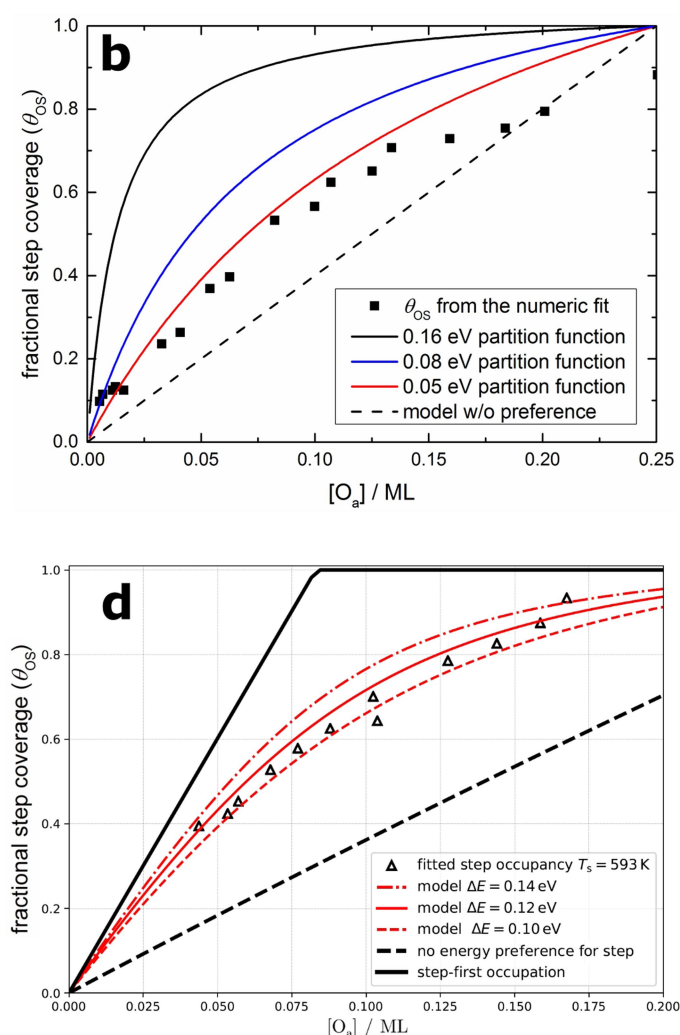
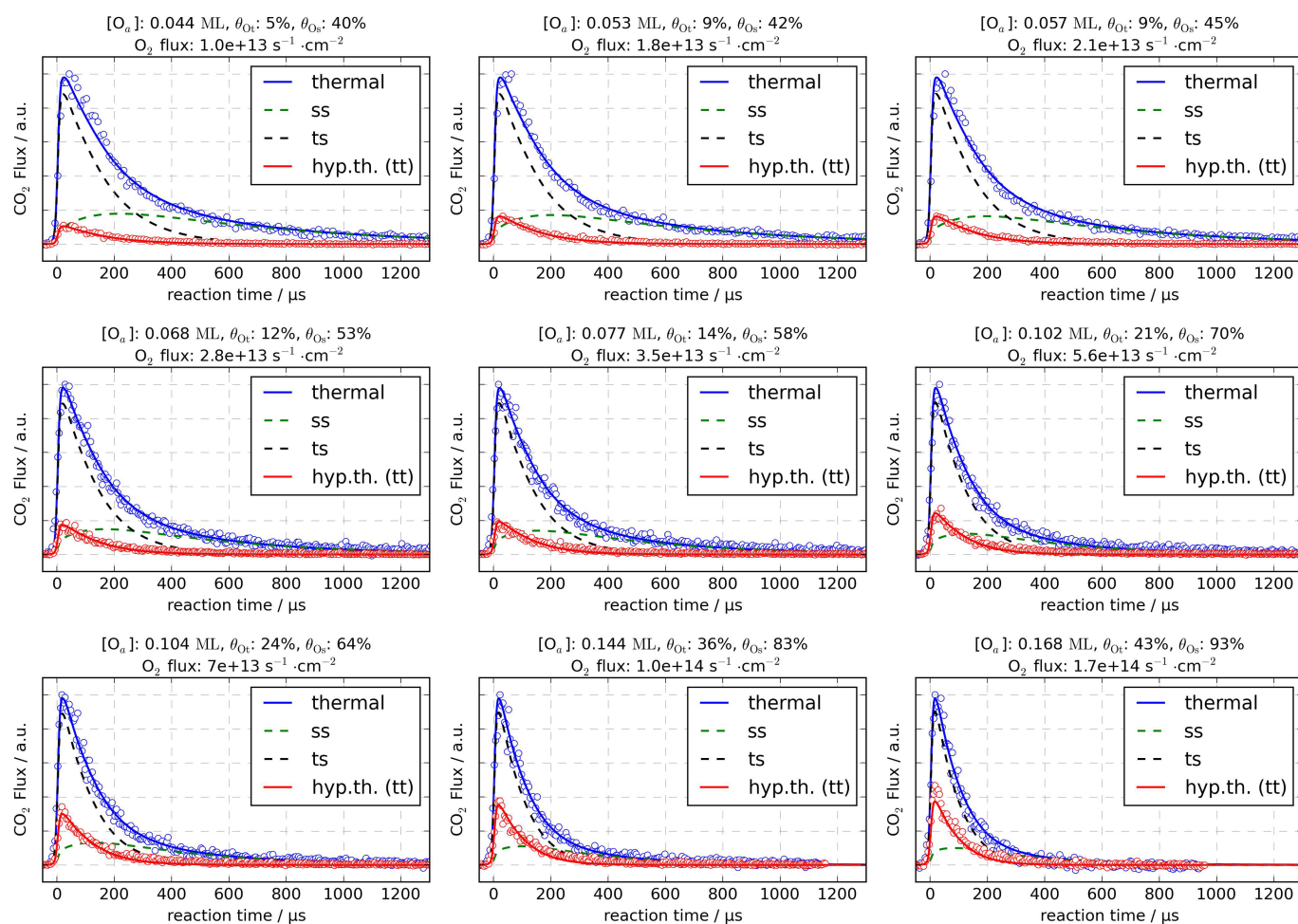


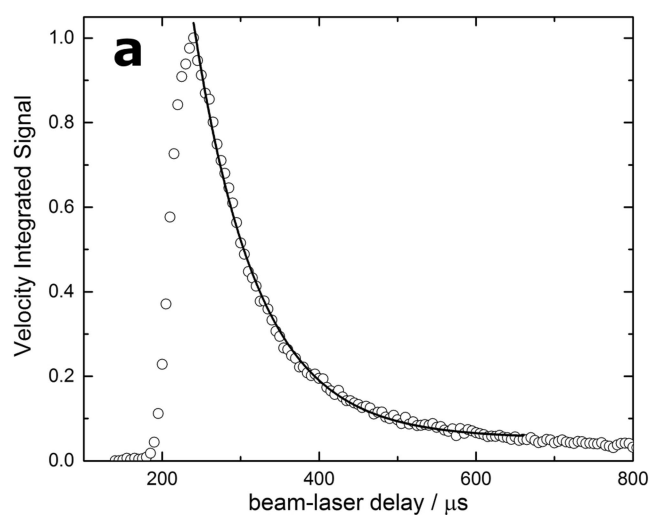
Fig. 6. All data are at  $340^\circ\text{C}$ . **c**, A comparison of values of  $[O_a]$  obtained from titrations on Pt(332) with  $[O_a]$  values obtained from the numerical kinetic model. Note that the titration results are found to be independent of surface temperature under our conditions. **d**, The fractional step coverage ( $\theta_{os}$ ) plotted against  $[O_a]$  for Pt(332). The red lines are results from the partition function calculated for different binding energy differences. The black triangles are the results from the numerical analysis of the kinetic model. Two extreme cases—no oxygen-atom binding preference for steps (dashed) and large oxygen-atom binding preference for steps (solid)—for the partition function are shown as black lines.



**Extended Data Fig. 8 | Examples of the kinetic model fit to experimentally derived kinetic traces for Pt(332).** All three reactive contributions are shown. Here  $T_s = 320^\circ C$  and the oxygen coverage varies from 0.044 to 0.168 monolayers (denoted as  $[O_a]$ ). The fractional coverage

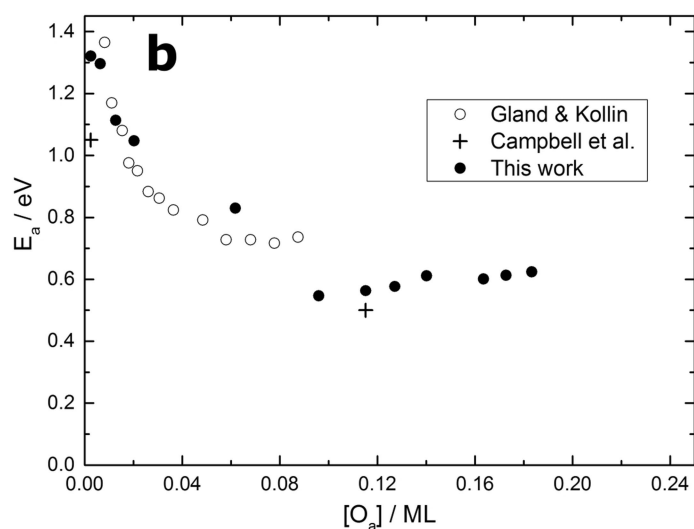
on terraces ( $\theta_{OT}$ ) and steps ( $\theta_{OS}$ ) is also indicated. The time-averaged O<sub>2</sub> flux is stated above each kinetic trace, the time-averaged CO flux was  $2.2 \times 10^{12} s^{-1} cm^{-2}$ .



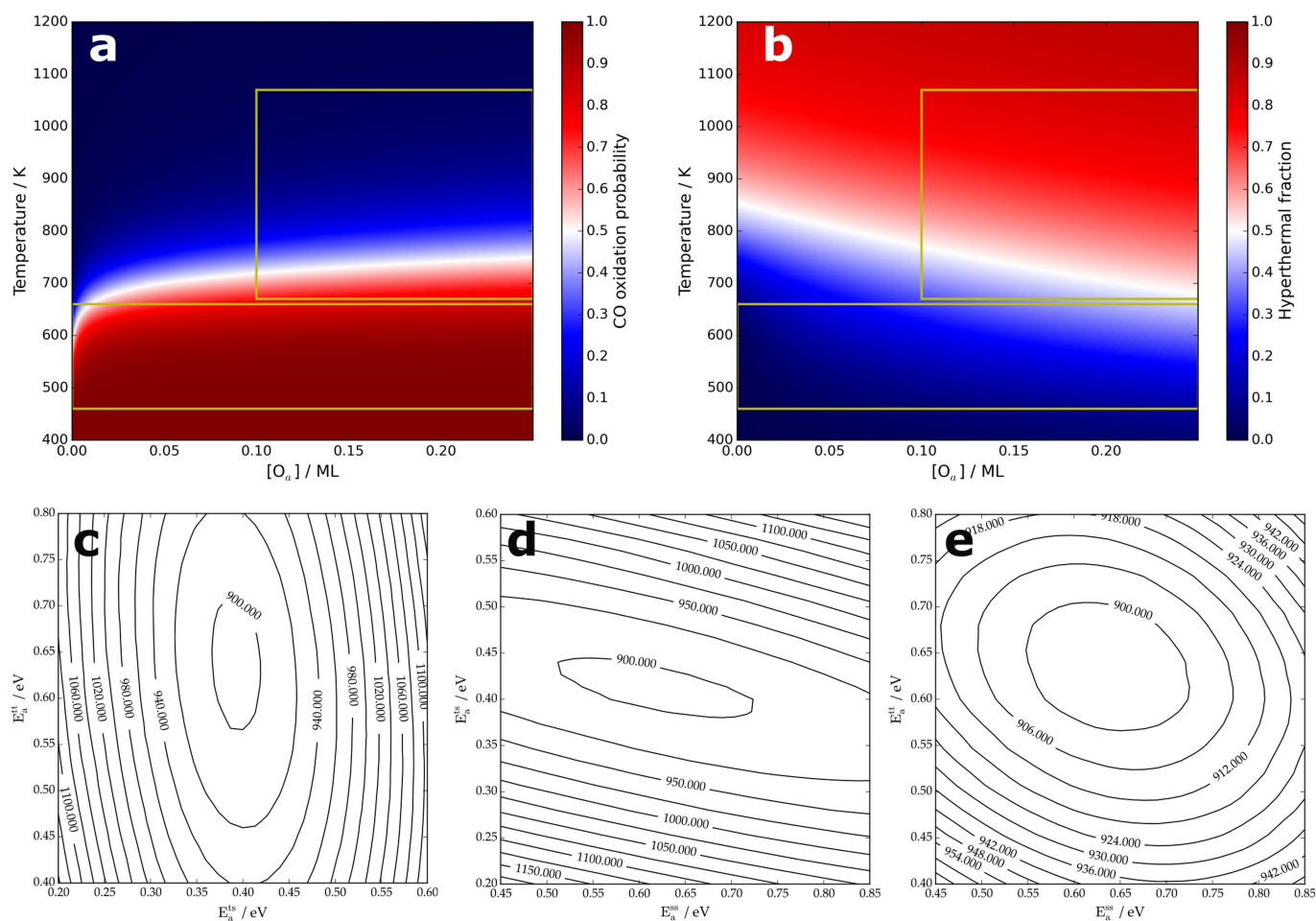


**Extended Data Fig. 9 | Activation energies of CO oxidation.**

**a**, Simulation of previous experiments. We show the kinetic trace integrated over velocity (hollow circles). The solid line shows an exponential fit to the simulated data. **b**, Previously reported activation



energies<sup>7,41</sup> (hollow circles) for CO oxidation that were based on product velocity unresolved measurements are compared to the results of this work when integrated over product velocity. See text for details.



**Extended Data Fig. 10 | Model predictions of CO oxidation on a Pt(111) crystal with a step density of 0.25%. a,** Total CO–CO<sub>2</sub> conversion efficiency as a function of temperature and oxygen coverage. The yellow boxes indicate (at low temperature) where past studies have been carried out and (at high temperatures) where industrial catalysts are used.

**b,** The relative importance of the hyperthermal (terrace) reaction as a function of temperature and oxygen coverage. **c–e,** Contour plots showing the total fit residual as a function of two activation energies. **y** against **x**: TT versus TS (**c**), TS versus SS (**d**), TT versus SS (**e**).

# Minimal East Antarctic Ice Sheet retreat onto land during the past eight million years

Jeremy D. Shakun<sup>1\*</sup>, Lee B. Corbett<sup>2</sup>, Paul R. Bierman<sup>2</sup>, Kristen Underwood<sup>3</sup>, Donna M. Rizzo<sup>3</sup>, Susan R. Zimmerman<sup>4</sup>, Marc W. Caffee<sup>5,6</sup>, Tim Naish<sup>7</sup>, Nicholas R. Golledge<sup>7</sup> & Carling C. Hay<sup>1</sup>

**The East Antarctic Ice Sheet (EAIS) is the largest potential contributor to sea-level rise. However, efforts to predict the future evolution of the EAIS are hindered by uncertainty in how it responded to past warm periods, for example, during the Pliocene epoch (5.3 to 2.6 million years ago), when atmospheric carbon dioxide concentrations were last higher than 400 parts per million. Geological evidence indicates that some marine-based portions of the EAIS and the West Antarctic Ice Sheet retreated during parts of the Pliocene<sup>1,2</sup>, but it remains unclear whether ice grounded above sea level also experienced retreat. This uncertainty persists because global sea-level estimates for the Pliocene have large uncertainties and cannot be used to rule out substantial terrestrial ice loss<sup>3</sup>, and also because direct geological evidence bearing on past ice retreat on land is lacking. Here we show that land-based sectors of the EAIS that drain into the Ross Sea have been stable throughout the past eight million years. We base this conclusion on the extremely low concentrations of cosmogenic <sup>10</sup>Be and <sup>26</sup>Al isotopes found in quartz sand extracted from a land-proximal marine sediment core. This sediment had been eroded from the continent, and its low levels of cosmogenic nuclides indicate that it experienced only minimal exposure to cosmic radiation, suggesting that the sediment source regions were covered in ice. These findings indicate that atmospheric warming during the past eight million years was insufficient to cause widespread or long-lasting meltback of the EAIS margin onto land. We suggest that variations in Antarctic ice volume in response to the range of global temperatures experienced over this period—up to 2–3 degrees Celsius above preindustrial temperatures<sup>4</sup>, corresponding to future scenarios involving carbon dioxide concentrations of between 400 and 500 parts per million—were instead driven mostly by the retreat of marine ice margins, in agreement with the latest models<sup>5,6</sup>.**

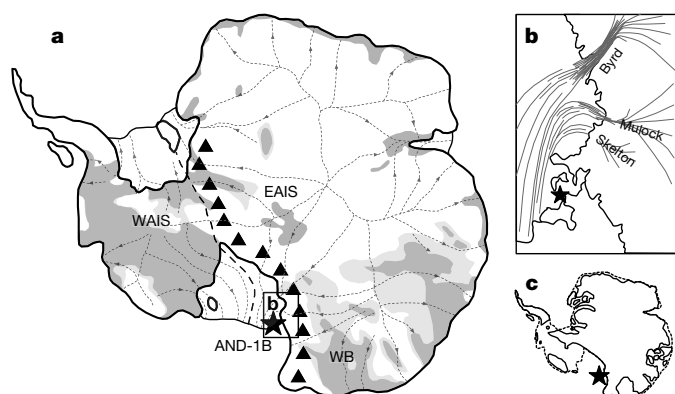
The configuration of the EAIS during the Pliocene epoch has been a longstanding topic of debate (see ref. <sup>7</sup> and references therein). Although the marine  $\delta^{18}\text{O}$  record (a proxy for global ice volume) is consistent with permanent EAIS establishment 14 million years ago (Ma), studies of Pliocene marine diatoms from glacial sediments high in the Transantarctic Mountains (TAMs) first raised the possibility that the ice sheet retreated substantially, resulting in an interior seaway, and then regrew as recently as 3 Ma. In apparent contradiction, geomorphic investigations identified ancient land surfaces in the TAMs, implying persistent polar desert conditions for the past several million years. This finding, together with evidence that the diatoms were not in situ but rather wind transported, led to doubts surrounding the scale of a late Neogene deglaciation. More recently, records from land-proximal marine sediments revealed sea surface temperatures that were much warmer than present, a lack of summer sea ice and the retreat of marine-based parts of the EAIS and the West Antarctic Ice Sheet (WAIS) numerous times during the Pliocene<sup>1,2,8</sup>.

Recent modelling may have reconciled these seemingly disparate observations, suggesting that the retreat of marine-based ice-sheet

sectors during the Pliocene allowed winds to loft diatoms from isostatically uplifted marine sediments onto TAM tills that may in fact be much older<sup>8,9</sup>. Atmospheric warming in these Pliocene simulations is insufficient to drive widespread surface melting and the retreat of land-terminating ice margins. Nonetheless, the resolution of the models is coarse relative to the width of potential ablation zones on the steep flanks of the ice sheet, and observations show that even in today's climate there is widespread surface melt as high as 1,300 m above sea level around Antarctica<sup>10</sup>. Moreover, Pliocene sea-level reconstructions cannot exclude the possibility that the EAIS margin retreated onto land. Global estimates of sea level during Pliocene interglacials show a wide range—from 5 m to more than 40 m higher than today<sup>3,11</sup>—owing to persistent uncertainties associated with extracting the sea-level signal from the marine  $\delta^{18}\text{O}$  record and from palaeo-shorelines. While most of this sea-level rise could be accounted for by loss of the present Greenland Ice Sheet (7 m) and/or marine-based portions of the WAIS (3 m) and EAIS (19 m), the highest estimates would also require input from EAIS land-based sectors (which contain 34 m sea-level equivalent)<sup>12</sup>. To date, however, there has been no direct evidence confirming that the EAIS margin retreated onto land during the past few million years.

Here we test directly for retreat of the EAIS margin onto land and for consequent exposure of the land surface to cosmic rays over the past 8 million years (Myr) by measuring concentrations of the cosmogenic nuclides <sup>10</sup>Be and <sup>26</sup>Al in quartz sands from the AND-1B sediment core, retrieved from beneath the Ross Ice Shelf (Figs. 1a and 2c). While diatom-rich units in AND-1B were used previously to identify numerous open-water intervals<sup>1</sup> associated with collapse of the WAIS<sup>8</sup>, we focused instead on terrigenous sediments from the intervening glacial units (Fig. 2) sourced from the EAIS. Cosmogenic nuclides are suitable for evaluations of past terrestrial retreat of the EAIS, because they are diagnostic of subaerial land exposure; these nuclides do not form in measurable concentrations under ice sheets, which shield rock surfaces from cosmic radiation. Cosmogenic nuclide concentrations in terrestrial materials from glaciated regions are controlled by a combination of ice-sheet extent and erosional processes<sup>13</sup>. Nuclide production is highest at the ground surface, decreases exponentially in the top few metres and then extends tens of metres deeper at much lower rates<sup>14</sup>. Exposure drives nuclide concentrations higher with time; however, nuclides are also lost owing to radioactive decay (the half-life of <sup>10</sup>Be is 1.39 Myr and that of <sup>26</sup>Al is 0.71 Myr) and erosion, which strips nuclide-rich surface material. Glaciation tends to favour both of these concentration-lowering processes because it halts nuclide production and, under areas of erosive ice, removes previously exposed material from the bed. The nuclide concentration preserved in sediments carried from land to an adjacent ocean reflects a convolution of exposure and burial history along with erosion depth, integrated along ice-flow lines and weighted by erosion rate<sup>13</sup>.

<sup>1</sup>Department of Earth and Environmental Sciences, Boston College, Chestnut Hill, MA, USA. <sup>2</sup>Department of Geology and Rubenstein School of the Environment and Natural Resources, University of Vermont, Burlington, VT, USA. <sup>3</sup>Civil and Environmental Engineering, University of Vermont, Burlington, VT, USA. <sup>4</sup>Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, CA, USA. <sup>5</sup>Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN, USA. <sup>6</sup>Department of Physics and Astronomy, Purdue University, West Lafayette, IN, USA. <sup>7</sup>Antarctic Research Centre, Victoria University of Wellington, Wellington, New Zealand. \*e-mail: jeremy.shakun@bc.edu

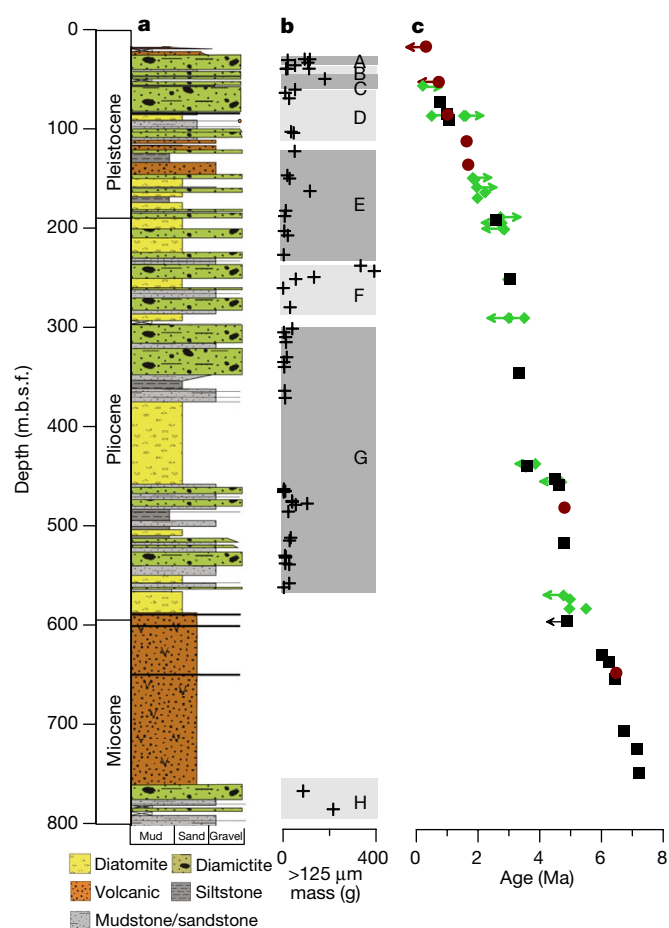


**Fig. 1 | Locations of the AND-1B core in the Ross Sea and ice-flow lines to the core site, along with simulated ice-sheet retreat during the Pliocene.** The location of the AND-1B core is marked with a star. **a**, Some areas upstream of the core site are below sea level at present<sup>12</sup> (light grey), but most would rise above sea level 10,000 years after deglaciation of marine-based ice-sheet sectors<sup>26</sup> (dark grey) and be exposed to cosmic radiation. WB, Wilkes Basin; triangles, Transantarctic Mountains. The dotted lines and arrows show ice-flow lines. Figure modified from ref. <sup>1</sup>. **b**, Ice flows to the core site from the Skelton and Mulock outlet glaciers of the EAIS at present, and probably did so throughout the Plio-Pleistocene<sup>15</sup>. Figure modified with permission from ref. <sup>15</sup>. **c**, Simulated Pliocene ice-sheet configuration (solid line), taken from ref. <sup>5</sup>, with retreat restricted almost entirely to marine-based sectors of the ice sheet. The dashed line shows the present ice extent.

The AND-1B sediments that we measured probably represent a relatively large area well upstream of the core site. Sand provenance indicators suggest that AND-1B glaciogenic sediments were likely to have been sourced from the Skelton and Mulock Glacier region of the TAMs throughout the Plio-Pleistocene<sup>15</sup>, and perhaps also the Wilkes Basin further inboard beneath the EAIS (Fig. 1). Indeed, Last Glacial Maximum tills further outboard in the Ross Sea appear to have been derived from the ice-sheet interior and from the entire width of the TAMs, which are traversed by outlet glaciers<sup>16</sup>. Modelling suggests that outlet glaciers may have been more active during past warm periods, sliding and eroding along much of their lengths<sup>17</sup>, and erosive zones would have migrated inward as the ice-sheet margin retreated (Extended Data Fig. 1). The volcanic rocks in McMurdo Sound surrounding the core site do not contain quartz<sup>15</sup>, precluding a local contribution. These source areas would have accumulated cosmogenic nuclides if substantial ice loss had occurred in the past. The TAMs are above sea level and would experience exposure to cosmic rays during ice-free periods. The Wilkes Basin lies below sea level today<sup>12</sup>, but much of it would have rebounded above sea level had the ice sheet experienced major retreat (Fig. 1a and Extended Data Fig. 2).

We evaluated whether AND-1B sediments have high concentrations of cosmogenic nuclides indicative of past ice retreat by using 62 diamictite samples from the top 786 metres of the core. These were amalgamated and purified to obtain sufficient quartz sand for measurement, yielding eight samples (designated A–H, from top to bottom), which span core intervals of 0.2–260.7 metres in thickness, or 0.002 Myr to 2.166 Myr in duration (Fig. 2; see also Methods and Supplementary Information). Procedural blanks extracted in addition to the samples quantify background nuclide abundances characteristic of sample processing and accelerator mass spectrometry analysis (see Methods and Supplementary Information).

In general, AND-1B nuclide concentrations are only marginally higher than those of procedural blanks, particularly for <sup>26</sup>Al (Fig. 3); thus, they are indicative of little, if any, near-surface exposure. We used several statistical techniques to evaluate the certainty with which sample measurements exceed analytical backgrounds. First, cumulative frequency distributions suggest that the sample population as a whole contains low but measurable concentrations of both <sup>10</sup>Be and <sup>26</sup>Al (Extended Data Fig. 3). Second, we used both frequentist and



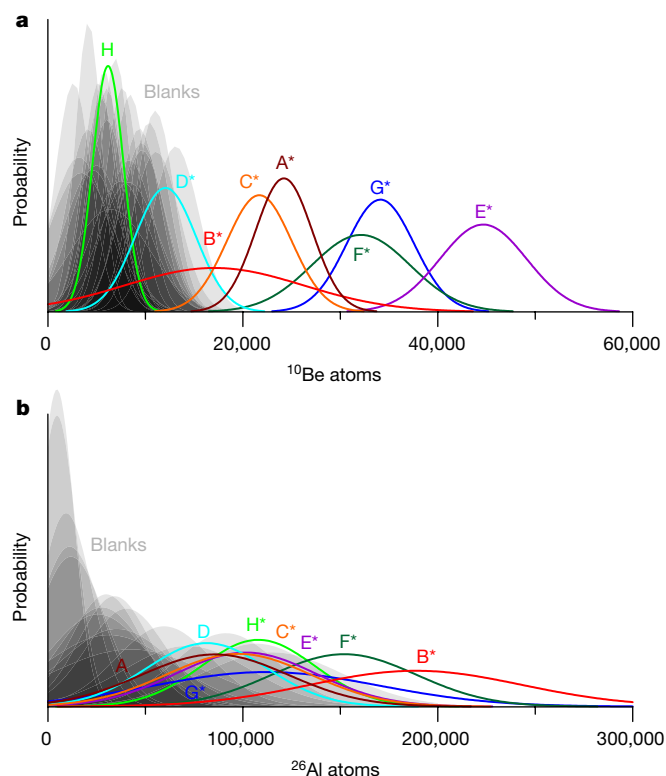
**Fig. 2 | AND-1B stratigraphy and cosmogenic nuclide samples.**

**a**, Lithostratigraphy of the AND-1B core<sup>27</sup> (m.b.s.f., metres below sea floor). **b**, Masses of sand (grains greater than 125 µm in size) in the 62 samples used here (plus symbols; see also Supplementary Information). Quartz was extracted from these samples and amalgamated, yielding eight final samples (labelled A to H and shaded grey) used for measurement of cosmogenic nuclides. We note that there is a roughly 200-m gap between samples G and H. **c**, Age control points based on palaeomagnetic events (black squares), <sup>40</sup>Ar/<sup>39</sup>Ar ratios (maroon circles) and diatom data (green diamonds)<sup>27</sup>. Arrows designate maximum or minimum limiting ages.

Bayesian statistical approaches to consider whether the overall sample population as well as individual samples contain cosmogenic nuclides at concentrations above background (Methods; Extended Data Tables 1 and 2). We found that all but the oldest sample contain more <sup>10</sup>Be than process blanks, and six of eight samples (B, C, E, F, G and H) contain <sup>26</sup>Al above blank level. We corrected sample nuclide concentrations for background using blank values, and for decay on the sea floor using the core-age model (see Methods). Although the <sup>26</sup>Al/<sup>10</sup>Be ratio can be used to monitor burial duration after exposure—because the faster decay of <sup>26</sup>Al relative to <sup>10</sup>Be causes the ratio to decrease from its production value<sup>14</sup>—the relatively large uncertainties of the low-concentration AND-1B data preclude resolution of meaningful decay-corrected <sup>26</sup>Al/<sup>10</sup>Be ratios.

The presence of <sup>10</sup>Be and <sup>26</sup>Al in AND-1B sediments provides unequivocal evidence for past landscape exposure in source regions. The relatively short half-life of <sup>26</sup>Al requires that this exposure occurred after EAIS expansion at 14 Ma, otherwise none would remain today (Extended Data Fig. 4). However, the decay-corrected nuclide concentrations are extremely low, declining from only about 12,000 to 600 <sup>10</sup>Be atoms per gram, and from 120,000 to 6,000 <sup>26</sup>Al atoms per gram, over the Plio-Pleistocene (Fig. 4a and Extended Data Fig. 4)—the equivalent of only 150–3,000 years of surface exposure at sea level (where the <sup>10</sup>Be production rate is about four atoms



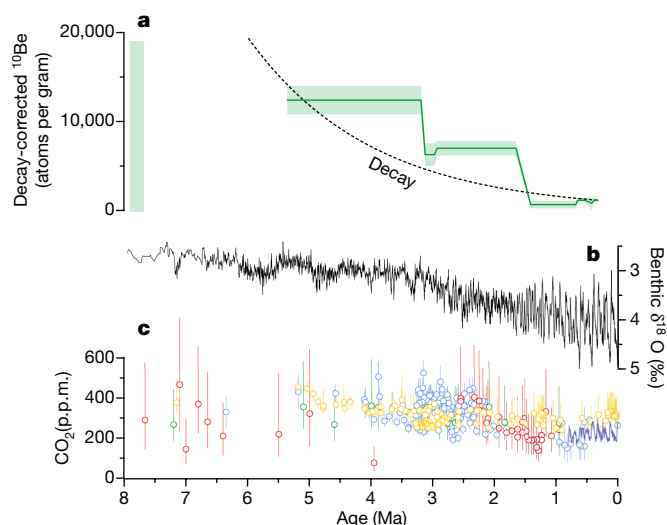


**Fig. 3 | Cosmogenic nuclide abundances.** **a**, **b**, Probability distribution functions showing the measured abundance (total number of atoms) of  $^{10}\text{Be}$  (**a**) and  $^{26}\text{Al}$  (**b**) in AND-1B samples (coloured lines). Also included are all laboratory process blanks run alongside low-level samples in the same hood by the same operator (shown as translucent grey). Labels A–H indicate samples from top to bottom in the core, and asterisks denote samples that have nuclide abundances above blank level at 90% confidence according to Bayesian analysis (see Methods). Note that the x axes in the two panels have different scales.

per gram per year), and even less at higher elevations. It is possible that these small concentrations of nuclides result from incorporation of minor amounts of material derived from nunataks in the TAMs<sup>18</sup>. In any case, these low concentrations are evidence that any Pliocene land-based ice-sheet retreat upstream of AND-1B was at most very limited in duration or extent. Several arguments support this interpretation.

First, modelling suggests that in most plausible scenarios, the AND-1B nuclide concentrations are consistent with spatially or temporally limited ice retreat during glacial cycles of the Pliocene and especially the Pleistocene (Methods); repeated exposures, even if relatively brief, would have led to higher concentrations than we measured (Extended Data Fig. 5). A single, long-lasting Pliocene exposure event—perhaps during the warm, roughly 3.6–3.4-Ma interval implied by a 60-m-thick diatomite in the AND-1B core<sup>1</sup>—could have yielded  $^{10}\text{Be}$  concentrations consistent with our Pliocene data if erosion rates were at least tens of metres per million years once the ice cover returned. However, without continued exposure, such high erosion rates cause  $^{10}\text{Be}$  concentrations to decline more steeply over time than observed in our record (Extended Data Figs. 6 and 7), and frequent exposure seems unlikely given that temperatures and sea level were rarely higher than at present during the Pleistocene<sup>19,20</sup>. The decreasing trend in AND-1B nuclide concentrations instead nearly follows a decay curve (Fig. 4a)—a trend that is most consistent with low rates of bedrock erosion under continuous ice cover since the last episodes of exposure.

Second, the persistent ice cover suggested by our record is also consistent with the perennially frozen conditions that some have argued are needed to explain the extreme landscape stability in the TAMs



**Fig. 4 | AND-1B  $^{10}\text{Be}$  record.** **a**, Decay-corrected  $^{10}\text{Be}$  concentrations from the AND-1B core. Shading shows  $1\sigma$  uncertainty. The vertical green bar near 8 Ma represents the possible range of decay-corrected concentrations in the oldest AND-1B sample that was below detection limit at the time of measurement (see Methods). The dashed black line shows the decay curve that would end at the youngest AND-1B sample. **b**, Benthic  $\delta^{18}\text{O}$  record, a proxy for global ice volume and deep ocean temperature<sup>28</sup>. **c**, Atmospheric  $\text{CO}_2$  concentrations from ice cores<sup>29</sup> (purple line towards the right), and from palaeosol  $\delta^{13}\text{C}$  (red), alkenone  $\delta^{13}\text{C}$  (yellow), stomata (green) and marine  $\delta^{11}\text{B}$  (blue), with  $1\sigma$  uncertainties<sup>30</sup>.

over the past few million years, which is reflected in high cosmogenic nuclide concentrations in exposed surfaces and intact surficial ash deposits<sup>7</sup>.

Finally, relative stability of the EAIS is evident when the AND-1B cosmogenic marine sediment record is compared with similar records from other ice sheets. Decay-corrected AND-1B  $^{10}\text{Be}$  concentrations are several times lower than values for Greenlandic sediment over the same time period<sup>13</sup> (Plio-Pleistocene averages of 7,000 and 25,000 atoms per gram, respectively), which are in turn several times lower than values for tills from the Laurentide Ice Sheet<sup>21,22</sup> (Pleistocene average of 70,000 atoms per gram). These patterns are consistent with repeated North American interglacial exposure sustaining higher nuclide concentrations, a generally present but dynamic Greenland Ice Sheet, and persistent Antarctic ice cover maintaining low  $^{10}\text{Be}$  concentrations.

How well does AND-1B represent the broader EAIS? Models simulating EAIS deglaciation suggest that ice may linger in the higher elevations of the TAMs, but all show the ice sheet generally contracting inwards, with considerable land area that is currently upstream of the core site being exposed relatively early during retreat<sup>23,24</sup>. Furthermore, these models suggest that warming consistent with even the most intense interglacials of the past 8 Myr would have produced minimal land exposure around the entire EAIS margin<sup>6,8</sup>. The near absence of cosmogenic nuclides in our record therefore rules out substantial and long-lasting EAIS retreat onto land during the past 8 Myr—a finding that is consistent with recent seismic imaging off the Aurora Basin<sup>25</sup>, and which could be confirmed by similar cosmogenic nuclide measurements in cores elsewhere around Antarctica. Our results also put an upper limit on estimates of Pliocene sea level, because melting of all marine-based ice in Antarctica<sup>12</sup> and the Greenland Ice Sheet could contribute at most 30 m of sea-level rise. Together with prior evidence for open waters at the AND-1B site during the Pliocene<sup>1</sup>, our findings agree with model simulations which show that the terrestrial EAIS experiences minimal melt when carbon dioxide levels are at their present value of roughly 400 parts per million for extended periods of time, whereas some marine-based ice-sheet sectors largely disappear<sup>5,23</sup> (Figs. 1c and 4c).

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0155-6>.

Received: 5 August 2017; Accepted: 20 March 2018;

Published online 13 June 2018.

- Naish, T. et al. Obliquity-paced Pliocene West Antarctic ice sheet oscillations. *Nature* **458**, 322–328 (2009).
- Cook, C. P. et al. Dynamic behaviour of the East Antarctic ice sheet during Pliocene warmth. *Nat. Geosci.* **6**, 765–769 (2013).
- Dutton, A. et al. Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science* **349**, aaa4019 (2015).
- Haywood, A. M., Dowsett, H. J. & Dolan, A. M. Integrating geological archives and climate models for the mid-Pliocene warm period. *Nat. Commun.* **7**, 10646 (2016).
- DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. *Nature* **531**, 591–597 (2016).
- Golledge, N. R. et al. Antarctic climate and ice-sheet configuration during the early Pliocene interglacial at 4.23 Ma. *Clim. Past* **13**, 959–975 (2017).
- Barrett, P. J. Resolving views on Antarctic Neogene glacial history—the Sirius debate. *Earth Env. Sci. Trans. R. Soc. Edinburgh* **104**, 31–53 (2013).
- Pollard, D. & DeConto, R. M. Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458**, 329–332 (2009).
- Scherer, R. P., DeConto, R. M., Pollard, D. & Alley, R. B. Windblown Pliocene diatoms and East Antarctic Ice Sheet retreat. *Nat. Commun.* **7**, 12957 (2016).
- Kingslake, J., Ely, J. C., Das, I. & Bell, R. E. Widespread movement of meltwater onto and across Antarctic ice shelves. *Nature* **544**, 349–352 (2017).
- Raymo, M., Mitrovica, J. X., O'Leary, M. J., DeConto, R. & Hearty, P. Departures from eustasy in Pliocene sea-level records. *Nat. Geosci.* **4**, 328–332 (2011).
- Fretwell, P. et al. Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *Cryosphere* **7**, 375–393 (2013).
- Bierman, P. R., Shakun, J. D., Corbett, L. B., Zimmerman, S. R. & Rood, D. H. A persistent and dynamic East Greenland Ice Sheet over the past 7.5 million years. *Nature* **540**, 256–260 (2016).
- Gosse, J. C. & Phillips, F. M. Terrestrial in situ cosmogenic nuclides: theory and application. *Quat. Sci. Rev.* **20**, 1475–1560 (2001).
- Talarico, F. M., McKay, R. M., Powell, R. D., Sandroni, S. & Naish, T. Late Cenozoic oscillations of Antarctic ice sheets revealed by provenance of basement clasts and grain detrital modes in ANDRILL core AND-1B. *Global Planet. Change* **96**, 23–40 (2012).
- Farmer, G. L. & Licht, K. J. Generation and fate of glacial sediments in the central Transantarctic Mountains based on radiogenic isotopes and implications for reconstructing past ice dynamics. *Quat. Sci. Rev.* **150**, 98–109 (2016).
- Golledge, N. R. & Levy, R. H. Geometry and dynamics of an East Antarctic Ice Sheet outlet glacier, under past and present climates. *J. Geophys. Res. Earth Surf.* **116**, F03025 (2011).
- Jones, R. S. et al. Cosmogenic nuclides constrain surface fluctuations of an East Antarctic outlet glacier since the Pliocene. *Earth Planet. Sci. Lett.* **480**, 75–86 (2017).
- Rohling, E. J. et al. Sea-level and deep-sea-temperature variability over the past 5.3 million years. *Nature* **508**, 477–482 (2014); corrigendum **510**, 432 (2014).
- Snyder, C. W. Evolution of global temperature over the past two million years. *Nature* **538**, 226–228 (2016).
- Balco, G., Stone, J. O. & Jennings, C. Dating Plio-Pleistocene glacial sediments using the cosmic-ray-produced radionuclides Be-10 and Al-26. *Am. J. Sci.* **305**, 1–41 (2005).
- Rovey, C. W. & Balco, G. Paleoclimatic interpretations of buried paleosols within the pre-Illinoian till sequence in northern Missouri, USA. *Palaeogeogr. Palaeoclim. Palaeoecol.* **417**, 44–56 (2015).
- Gasson, E., DeConto, R. M., Pollard, D. & Levy, R. H. Dynamic Antarctic ice sheet during the early to mid-Miocene. *Proc. Natl Acad. Sci. USA* **113**, 3459–3464 (2016).
- Winkelmann, R., Levermann, A., Ridgwell, A. & Caldeira, K. Combustion of available fossil fuel resources sufficient to eliminate the Antarctic Ice Sheet. *Sci. Adv.* **1**, e1500589 (2015).
- Gulick, S. P. S. et al. Initiation and long-term instability of the East Antarctic Ice Sheet. *Nature* **552**, 225–229 (2017).
- Hay, C. et al. The sea-level fingerprints of ice-sheet collapse during interglacial periods. *Quat. Sci. Rev.* **87**, 60–69 (2014).
- Wilson, G. S. et al. Neogene tectonic and climatic evolution of the Western Ross Sea, Antarctica—chronology of events from the AND-1B drill hole. *Global Planet. Change* **96**, 189–203 (2012).
- Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
- Lüthi, D. et al. High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* **453**, 379–382 (2008).
- Foster, G. L., Royer, D. L. & Lunt, D. J. Future climate forcing potentially without precedent in the last 420 million years. *Nature Comm.* **8**, 14845 (2017).

**Acknowledgements** We thank the Antarctic Research Facility for AND-1B samples, and J. X. Mitrovica for his help in performing the glacial isostatic adjustment modelling. This research was supported by National Science Foundation (NSF) grant ARC-1023191 (to P.R.B. and L.B.C.); Boston College start-up funds (to J.D.S.); Vermont Established Program to Stimulate Competitive Research (EPSCoR) grants EPS-1101317 and NSF OIA 1556770 (to K.U. and D.M.R.); NSF grant EAR-1153689 (to M.W.C.); and the New Zealand Ministry of Business Innovation and Employment contract C05X1001 (to T.N. and N.R.G.). This is Lawrence Livermore National Laboratory project LLNL-JRNL-735619.

**Reviewer information** Nature thanks J. Gosse, E. Gasson, J. Willenbring and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** J.D.S. and P.R.B. conceived the study. L.B.C. performed laboratory work. K.U. and D.M.R. conducted statistical analyses. S.R.Z. and M.W.C. made isotopic measurements. T.N. and N.R.G. contributed to data interpretation. C.C.H. performed glacial isostatic adjustment simulations. All authors contributed to the preparation of the manuscript.

**Competing interests** The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0155-6>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0155-6>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to J.D.S.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

**Sediment samples.** We prepared a total of 62 samples, ranging in thickness from 2 cm to 70 cm, from the top 786 m of the AND-1B core. The top 61 samples are from Facies 10, consisting of diamicton interpreted as subglacial sediments, although possibly also associated with rainout from floating ice and mass flows<sup>31</sup>. The lowest sample is from Facies 4, interpreted as hemipelagic muds and clasts rained out from icebergs or an ice shelf<sup>31</sup>. The samples were disaggregated by soaking in weak acid and then wet-sieved; then, because of their small masses, fractions of greater than 125  $\mu\text{m}$  from adjacent samples were amalgamated until they totalled around 200 g, yielding 14 samples. Quartz was isolated through an initial heated hydrochloric acid ultrasonic etching, followed by repeated weak (0.25%) hydrofluoric acid/nitric acid etching; the fraction that was greater than 1,000  $\mu\text{m}$  was sieved to remove larger grains freed up during acid disaggregation. Because some of the resulting quartz separates were too small for  $^{10}\text{Be}$  measurement, we amalgamated adjacent samples once again, producing eight final samples with quartz masses of 14–20 g, spanning core intervals of 0.2–260.7 m in thickness, or 0.002 Myr to 1.763 Myr in duration. Sample processing details are given in Supplementary Information.

**Cosmogenic nuclide extraction.**  $^{10}\text{Be}$  and  $^{26}\text{Al}$  were extracted from quartz at the University of Vermont, following the methods of ref. <sup>32</sup>. We added around 250  $\mu\text{g}$  of  $^9\text{Be}$  to each sample using a beryl carrier made at the University of Vermont. The total  $^{27}\text{Al}$  was quantified using inductively coupled plasma optical emission spectrometry analysis of replicate aliquots removed from samples directly after digestion; measurements were made using an internal standard (Y) and two different Al emission lines. The  $^{10}\text{Be}/^9\text{Be}$  ratios were measured at the Lawrence Livermore National Laboratory and normalized to primary standard 07KNSTD3110, assuming a  $^{10}\text{Be}/^9\text{Be}$  ratio of  $2.85 \times 10^{-12}$  (ref. <sup>33</sup>).  $^{26}\text{Al}/^{27}\text{Al}$  ratios were measured at the Purdue Rare Isotope Measurement Laboratory and normalized to KNSTD standard 26Al-1-05-2, with an assumed  $^{26}\text{Al}/^{27}\text{Al}$  ratio of  $1.82 \times 10^{-12}$  (ref. <sup>34</sup>). Isotopic data are given in Supplementary Information.

**Blanks.** Two procedural blanks were prepared with the samples during nuclide extraction to estimate the background nuclide abundances resulting from both laboratory sample processing and accelerator mass spectrometry (AMS) analysis. To characterize more fully the possible range of backgrounds, we also considered all blanks run by the same operator in the same fume hood (dedicated exclusively to the preparation of low-ratio samples) with other long-buried samples from beneath the Greenland Ice Sheet ( $n = 26$  for  $^{10}\text{Be}$ ;  $n = 18$  for  $^{26}\text{Al}$ ). We quantified the central tendency and dispersion of the sample and blank populations using their arithmetic means and standard errors. All blanks were measured at the same AMS facilities as the samples (Lawrence Livermore National Laboratory for  $^{10}\text{Be}$ , Purdue Rare Isotope Measurement Laboratory for  $^{26}\text{Al}$ ).

**Statistical analysis.** Traditionally, frequentist methods (that is, null-hypothesis significance tests; for example, the  $t$ -test) are applied to sample results and process blanks in order to determine whether the two groups are statistically different or if individual samples are statistically distinct from a group of blanks. A null hypothesis is framed (for example, 'the true mean of the samples is less than or equal to the true mean of the blanks') and is rejected for  $P$  values less than the significance level ( $\alpha$ ). However, under a frequentist approach, one cannot express support for the research hypothesis (for example, that a given sample result is greater than the mean of the blanks). A low  $P$  value simply means there is a low chance that our data would support the null hypothesis simply by random chance<sup>35</sup>. In other words, the probability of incorrectly rejecting the null (type I error) is low. On the other hand, a Bayesian analysis can support either a null or a research hypothesis, provide explicit information about the precision of parameter estimations in the form of credible intervals, and allow data to be examined in the context of the effect size<sup>35–37</sup>. Bayesian frameworks have the added advantage of allowing for non-normal distribution of datasets<sup>38</sup>, which can be a common occurrence when working with low-level detection samples<sup>39</sup>. Notably, our AND-1B datasets exhibit some non-normality. While  $^{10}\text{Be}$  blanks and samples and  $^{26}\text{Al}$  blanks were each from a normal distribution, as confirmed by the Shapiro–Wilks test (at a significance level,  $\alpha$ , of 0.05), the group of  $^{26}\text{Al}$  blanks was not normally distributed.

We applied Bayesian inference to estimate distribution parameters (including the mean and standard deviation) for our AND-1B datasets, and to compare parameters between the samples and blanks. In the context of the  $t$ -test, our response variables ( $^{10}\text{Be}$  and  $^{26}\text{Al}$  reported in the samples and the blanks) were modelled as having a  $t$ -distribution to accommodate potential outliers. The  $t$ -distribution is similar to a normal distribution, parameterized by the mean ( $\mu$ ) and the standard deviation ( $\sigma$ ), but also includes a third parameter ( $\nu$ ) that accommodates heavy-tailed distributions ( $\nu < 30$ ) and near-normal distributions ( $\nu = 30$ ).

Model parameters included  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$ ,  $\sigma_2$  and  $\nu$ , with the subscripts denoting the identity of the blanks (group 1) and samples (group 2). Non-informative priors were assigned to each parameter. A normal distribution was assigned to  $\mu_1$  and  $\mu_2$ , centred on the empirical mean of pooled data (both blank and sample groups), with a variance ranging from 0.2 to 5 times the standard deviation of the pooled data. Standard deviation parameters— $\sigma_1$  and  $\sigma_2$ —were each assigned a prior distributed

as a gamma density with shape and rate parameters "derived from the desired mode and standard deviation of the gamma distribution"<sup>40</sup>. Equal normality was assumed for both groups, with a prior for  $\nu$  distributed as a gamma density with shape and rate parameters "derived from the desired mean and standard deviation of the gamma distribution"<sup>40</sup>.

We used Gibbs sampling to obtain samples from the posterior distribution and to estimate the mean, standard deviation, standard error of the mean, mode, quantiles and credible intervals for each Bayesian model parameter. Markov-chain Monte Carlo sampling was implemented in R code<sup>41</sup> using the 'BEST' package<sup>42</sup>, which relies on 'JAGS'<sup>43</sup> and 'coda'<sup>44</sup> code. Sampling was conducted with three chains, for 100,000 iterations with a thinning factor of 1, after discarding the initial 4,500 iterations for adaptation and burn-in phases. Convergence was confirmed by visual examination of trace plots and the Gelman–Rubin statistic<sup>45</sup> (that is, requiring a potential shrink-reduction factor of less than 1.1).

We first compared the blanks with the samples for each nuclide (expressed in atoms), applying a one-sided, two-group  $t$ -test (2GTT) using both frequentist and Bayesian approaches. Notably, our blank results included all laboratory process blanks run alongside low-level samples in the same hood by the same operator over time ( $n = 26$  for  $^{10}\text{Be}$  and  $n = 18$  for  $^{26}\text{Al}$ ).

A frequentist one-sided 2GTT (unequal variances, Welch approximation) determined that the mean of the samples may be greater than the mean of the blanks for both  $^{10}\text{Be}$  and  $^{26}\text{Al}$ . The null hypothesis (that the true mean of the samples is less than or equal to the true mean of the blanks) was rejected at  $\alpha = 0.05$  ( $P < 0.05$ ; Extended Data Table 1).

A Bayesian, one-sided 2GTT yielded consistent results, but provided support for the research hypothesis (that is, that the sample mean is greater than the blank mean) and quantified the probability. Results indicate that the two groups (samples and blanks) are credibly different for both  $^{10}\text{Be}$  and  $^{26}\text{Al}$ ; moreover, there is a greater than 99% probability that the sample mean is greater than the blank mean for each nuclide (Extended Data Table 1). The 95% credible interval on the estimated difference between group means excludes zero for both  $^{10}\text{Be}$  and  $^{26}\text{Al}$ .

Next, we compared each individual sample with the mean of the blanks to determine whether the samples contained credible quantities of  $^{10}\text{Be}$  and  $^{26}\text{Al}$ , using both frequentist and Bayesian methods. The two methods yielded contrasting outcomes, with the value of Bayesian over frequentist methods becoming especially apparent from this analysis.

We performed the frequentist 1GTT to evaluate the null hypothesis that the sample result was less than or equal to the blank mean. For  $^{10}\text{Be}$ , we obtained a low  $P$  value (less than 0.10) for seven of the eight samples (A to G), rejecting the null hypothesis. However, this 1GTT failed to reject the null hypothesis for reported  $^{10}\text{Be}$  in sample H. For  $^{26}\text{Al}$ , a low  $P$  value (less than 0.10) resulted in a rejection of the null hypothesis for all eight samples. However, frequentist analysis does not permit us to express support for the research hypothesis—that is, that  $^{26}\text{Al}$  is credibly present in these eight samples, or that  $^{10}\text{Be}$  is credibly present in seven of the eight samples—only that there is a low probability of a type I error.

Bayesian 1GTTs were carried out by defining a region of uncertainty (ROU) around the sample-specific estimate equal to  $\pm 1\sigma$ . Bayesian 1GTT results indicated that  $^{10}\text{Be}$  is credibly present above the mean blank value in seven of the eight samples (all but H). In each credible case, the blank mean is less than the sample value, and a 90% credibility interval on the posterior distribution of the mean of the blanks fully excludes the  $\pm 1\sigma$  ROU surrounding the sample value (Extended Data Fig. 8a). By contrast, for sample H, the mean of the blanks is greater than the reported sample value (Extended Data Fig. 8b). An estimated 92.9% of the posterior distribution is above the sample value, and 89.9% of the posterior distribution of the mean is located within the sample-specific ROU (Extended Data Table 2).

In the case of  $^{26}\text{Al}$ , Bayesian 1GTT results indicated that this constituent is credibly present above the mean blank value in samples B, C, E, F, G and H only. Samples A and D may be, to varying degrees, indistinguishable from the process blanks (Extended Data Table 2). Although the 90% credibility interval on the posterior distribution of the blank mean was less than, and fully excluded, the respective sample value, a percentage of the posterior distribution (17.5% and 10.8%, respectively, for A and D) was located within the sample-specific ROU—including portions of the 90% credibility interval (Extended Data Fig. 8c).

Thus, the Bayesian 1GTT results contrast with those of the frequentist null-hypothesis significance tests. Although significant  $P$  values (at  $\alpha = 0.10$ ) obtained in the frequentist 1GTT may have led one to conclude that  $^{10}\text{Be}$  is present in seven out of eight samples and that  $^{26}\text{Al}$  is present in all eight samples with a 'confidence' interval of 90%, we can only state a low chance of having incorrectly rejected a true null hypothesis (that is, generated a 'false positive') for these samples. The Bayesian approach, however, allowed us to express support for the research hypothesis that  $^{10}\text{Be}$  is credibly present above the mean of the blanks in samples A to G, and that  $^{26}\text{Al}$  is credibly present above background only in samples B, C, E, F, G and H.

It is difficult to explain the presence of  $^{26}\text{Al}$  but the absence of  $^{10}\text{Be}$  in sample H given the shorter half-life of  $^{26}\text{Al}$ . Because the  $^{26}\text{Al}$  abundance in this sample barely



exceeds the threshold for statistical significance (Fig. 3b), we consider it dubious that this sample in fact contains measurable  $^{26}\text{Al}$ , and disregard it.

**Blank correction.** To estimate the concentration of nuclides in the samples, we subtracted the mean nuclide abundance of all the blanks. We quantified the blank uncertainty as the standard error of the mean, and combined it with the measurement uncertainty of the sample nuclide abundances in quadrature.

**Decay correction.** Using the age model from ref. <sup>27</sup>, we corrected measured cosmogenic nuclide concentrations for radioactive decay on the seafloor following deposition; this age model is based on palaeomagnetic events, biostratigraphic data and  $^{40}\text{Ar}/^{39}\text{Ar}$  ages. Because measurements were made on amalgamated subsamples that span considerable depth (and thus age) ranges in the core, we used the average age of the subsamples, weighted by their sand (more than  $125\text{ }\mu\text{m}$ ) masses. Our cosmogenic record is relatively insensitive to age-model uncertainties because the age uncertainties are substantially smaller than the half-lives of  $^{10}\text{Be}$  and  $^{26}\text{Al}$  (1.39 Myr, ref. <sup>46</sup>; and 0.71 Myr, ref. <sup>47</sup>).

**Limiting initial nuclide concentrations of blank-level samples.** We estimated the maximum concentrations that samples without measurable nuclides could have had when they were deposited and still decayed to blank levels by today. For example, our average  $^{10}\text{Be}$  blank has about 7,000 atoms. This implies that the oldest sample (H), which does not have measurable  $^{10}\text{Be}$ , could have had up to some 380,000 atoms (or around 19,000 atoms per gram for this 20.1-g sample) when it was deposited 8 Ma (5.75 half-lives ago).

**Antarctic erosion rates.** The AND-1B cosmogenic nuclide record is erosion weighted because sediments are sourced from areas in proportion to their erosion rate. In addition, areas of more rapid erosion contribute deeper-sourced, and thus nuclide-poorer, material, potentially diluting the cosmogenic nuclide concentration resulting from a past exposure event. Modelling, sediment backstacking techniques, subglacial imaging, and thermochronometry suggest that while Antarctic subglacial erosion rates may have reached up to 200 m per Myr in spatially isolated regions in the past, spatial averages were typically much lower ( $1\text{--}2\text{ m per Myr}$ )<sup>48–50</sup>, particularly during the late Cenozoic, when landscape dissection probably slowed with colder conditions and less aggressive glacial erosion<sup>49,51,52</sup>.

**Modelling hypothetical exposure scenarios.** We generated synthetic  $^{10}\text{Be}$  and  $^{26}\text{Al}$  records to inform our interpretation of the AND-1B record, using the MATLAB implementation of refs. <sup>53,54</sup>, and including nuclide production by muons. In a first experiment, we evaluated whether the  $^{26}\text{Al}$  measured in our record could have been produced before EAIS expansion at 14 Ma; we found that it could not. For instance, even in an extreme scenario with a high-elevation, non-eroding surface saturated with  $^{26}\text{Al}$  (namely, equal rates of nuclide production and loss by radioactive decay) at 14 Ma and subsequently decaying for 20 half-lives under cold-based, non-erosive ice cover, there would not be any detectable  $^{26}\text{Al}$  remaining today (Extended Data Fig. 4).

In a second set of simulations, we calculated  $^{10}\text{Be}$  and  $^{26}\text{Al}$  concentrations in sediment eroded from a bedrock profile averaged over multiple glacial cycles, driven by high-latitude production rates at sea level and at 2,000 m above sea level, and given various durations of exposure per cycle (0%–100%) and glacial erosion rates (0–100 m per Myr). AND-1B Pliocene  $^{10}\text{Be}$  concentrations are consistent with exposure lasting more than 10% of a glacial cycle only in cases with low-elevation sediment sources and very high erosion rates; most scenarios that are in agreement with our AND-1B data instead imply little to no exposure during the Pliocene, and especially during the Pleistocene (Extended Data Fig. 5).

Finally, we tested whether a single episode of exposure for 10 thousand years (kyr), 50 kyr, 100 kyr or 200 kyr during the mid-Pliocene could yield a  $^{10}\text{Be}$  record similar to ours. We assumed that the bedrock had no cosmogenic nuclides initially, that rock was eroded at various rates (0, 20 or 100 m per Myr), and that the resulting sediment was then transported instantaneously to the sea floor, where the radionuclides were subject to decay only. The resulting time series were degraded to the resolution of our isotopic record. The results of these simulations suggest that both short-exposure, low-erosion and long-exposure, high-erosion scenarios could yield  $^{10}\text{Be}$  concentrations similar to those of our AND-1B Pliocene samples, but that the rapid erosion in the latter scenarios decreases  $^{10}\text{Be}$  concentrations more quickly than we observe during the Pleistocene (Extended Data Fig. 6). The long mid-Pliocene exposure scenario would require either a substantial decrease in erosion rates from the Pliocene to the Pleistocene, or continued brief exposure events during the Pleistocene to match the entire AND-1B record. We also repeated the 200-kyr mid-Pliocene exposure simulation, but first mixed eroded bedrock into a deformable till layer before fluxing it to the ocean. We used deformable bed thicknesses of 1 m and 10 m and erosion rates of 10 m per Myr

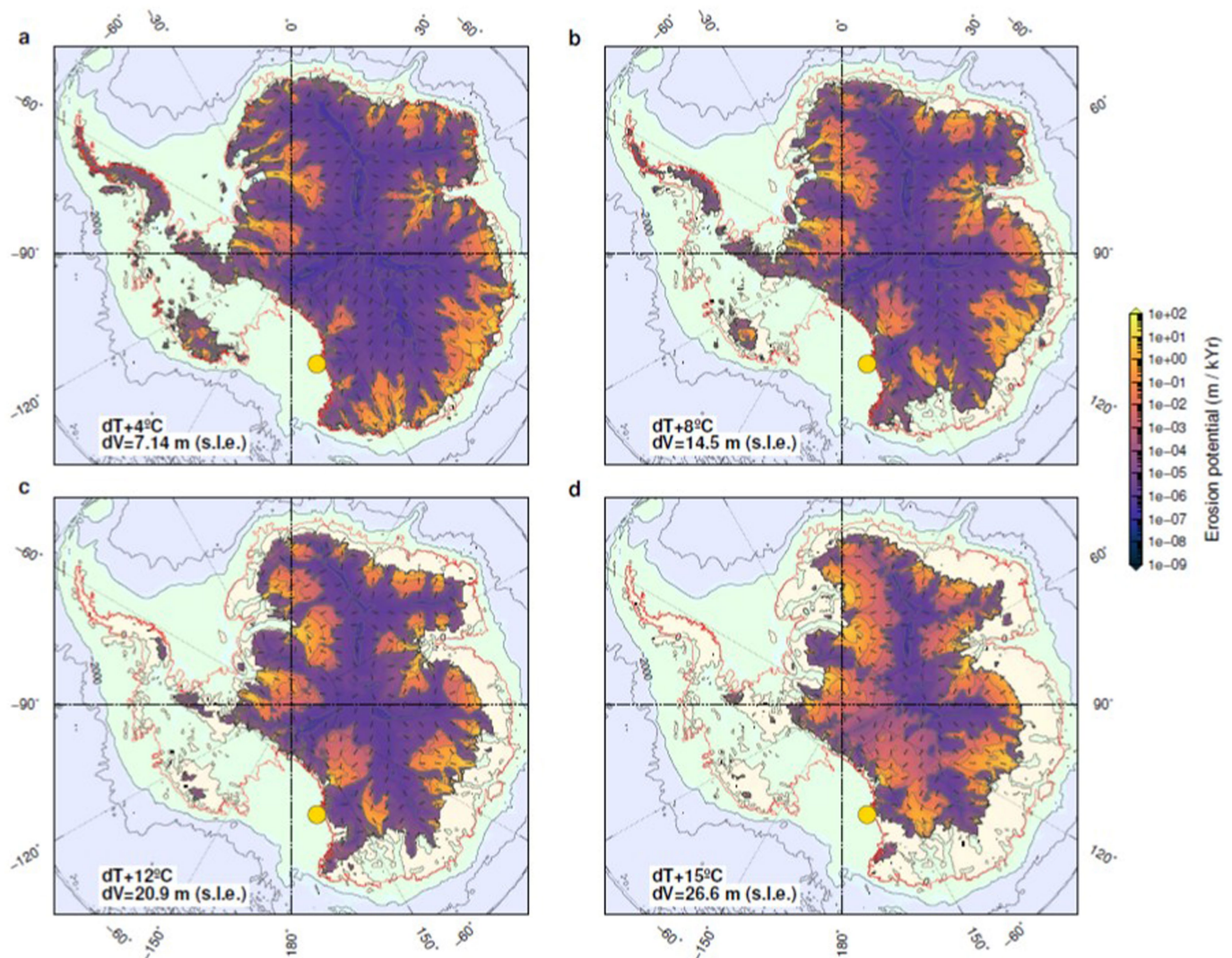
and 80 m per Myr, assumed instantaneous homogenization of eroded bedrock material throughout the deformable bed in each time step, and removed an equal amount of bed material to the ocean to keep the bed thickness constant. This sediment mixing reduces the magnitude of the mid-Pliocene exposure signal and extends its longevity through time in comparison with the bedrock-only simulations, but it similarly fails to replicate the AND-1B record; high erosion rates are needed to reduce Pliocene  $^{10}\text{Be}$  concentrations in the regolith to AND-1B values, but low erosion rates are needed to have slowly discharged the regolith to the ocean over the past few million years (Extended Data Fig. 7). More complicated scenarios would be required to fit the AND-1B record, such as long-lived pockets of higher- $^{10}\text{Be}$  regolith mixing with more rapidly eroding, and thus lower- $^{10}\text{Be}$ , bedrock.

**Data availability.** All AND-1B data generated here (sediment processing, sample and blank isotopic ratios and concentrations) are provided as Supplementary Information.

**Code availability.** The codes used to model cosmogenic nuclide exposure and the erosion scenarios shown in the Extended Data Figures are available at <https://github.com/shakun/Shakun-et-al-2018-Nature>.

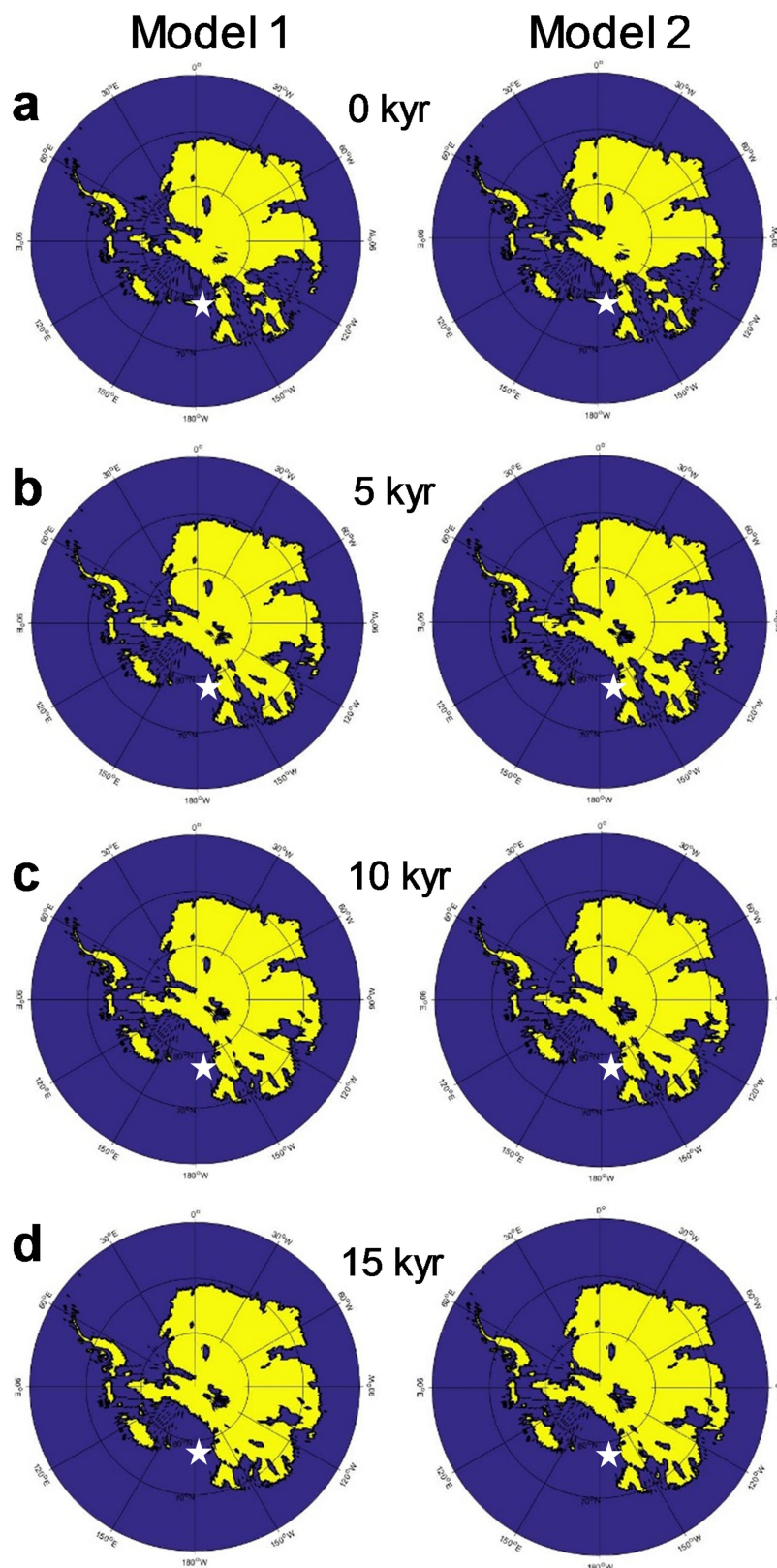
31. Krissek, L. et al. Sedimentology and stratigraphy of the AND-1B core, ANDRILL McMurdo Ice Shelf Project, Antarctica. *Terra Antarctica* **14**, 185–222 (2007).
32. Corbett, L. B., Bierman, P. R. & Rood, D. H. An approach for optimizing in situ cosmogenic  $^{10}\text{Be}$  sample preparation. *Quat. Geochronol.* **33**, 24–34 (2016).
33. Nishiizumi, K. et al. Absolute calibration of  $^{10}\text{Be}$  AMS standards. *Nucl. Instrum. Methods Phys. Res. B* **258**, 403–413 (2007).
34. Nishiizumi, K. Preparation of  $^{26}\text{Al}$  AMS standards. *Nucl. Instrum. Methods Phys. Res. B* **223–224**, 388–392 (2004).
35. Nuzzo, R. Statistical errors. *Nature* **506**, 150–152 (2014).
36. Kruschke, J. K. Bayesian estimation supersedes the t test. *J. Exp. Psychol.* **142**, 573–603 (2013).
37. Kruschke, J. K. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan* (Elsevier, London, 2015).
38. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman & Hall/CRC, London, 2004).
39. Currie, L. A. The measurement of environmental levels of rare gas nuclides and the treatment of very low-level counting data. *IEEE Trans. Nucl. Sci.* **19**, 119–126 (1972).
40. Kruschke, J. K. *Informed priors for Bayesian comparison of two groups* <http://doingbayesiandataanalysis.blogspot.com/2015/04/informed-priors-for-bayesian-comparison.html> (2015).
41. R Core Development Team. *R: a language and environment for statistical computing* <http://www.R-project.org/> (2016).
42. Kruschke, J. K. & Meredith, M. *BEST: Bayesian estimation supersedes the t-test* <https://cran.r-project.org/web/packages/BEST/index.html> (2015).
43. Plummer, M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd Int. Workshop on Distributed Statistical Computing* (eds Hornik, K. et al.) (2003).
44. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11 (2006).
45. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
46. Korschinek, G. et al. A new value for the half-life of  $^{10}\text{Be}$  by heavy-ion elastic recoil detection and liquid scintillation counting. *Nucl. Instrum. Methods Phys. Res. B* **268**, 187–191 (2010).
47. Norris, T. L., Gancarz, A. J., Rokop, D. J. & Thomas, K. W. Half-life of  $^{26}\text{Al}$ . *J. Geophys. Res. Solid Earth* **88**, B331–B333 (1983).
48. Jamieson, S. S. R., Sugden, D. E. & Hulton, N. R. J. The evolution of the subglacial landscape of Antarctica. *Earth Planet. Sci. Lett.* **293**, 1–27 (2010).
49. Thomson, S. N., Reinert, P. W., Hemming, S. R. & Gehrels, G. E. The contribution of glacial erosion to shaping the hidden landscape of East Antarctica. *Nat. Geosci.* **6**, 203–207 (2013).
50. Wellman, P. & Tingey, R. J. Glaciation, erosion and uplift over part of East Antarctica. *Nature* **291**, 142–144 (1981).
51. Bo, S. et al. The Gamburtsev mountains and the origin and early evolution of the Antarctic Ice Sheet. *Nature* **459**, 690–693 (2009).
52. Young, D. A. et al. A dynamic early East Antarctic Ice Sheet suggested by ice-covered fjord landscapes. *Nature* **474**, 72–75 (2011).
53. Heisinger, B. et al. Production of selected cosmogenic radionuclides by muons. *Geochim. Cosmochim. Acta* **66**, A558 (2002).
54. Balco, G., Stone, J. O., Lifton, N. A. & Dunai, T. J. A complete and easily accessible means of calculating surface exposure ages or erosion rates from  $^{10}\text{Be}$  and  $^{26}\text{Al}$  measurements. *Quat. Geochronol.* **3**, 174–195 (2008).
55. Gollledge, N. R., Levy, R. H., McKay, R. M. & Naish, T. R. East Antarctic ice sheet most vulnerable to Weddell Sea warming. *Geophys. Res. Lett.* **44**, 2343–2351 (2017).
56. Peltier, W. R. Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci.* **32**, 111–149 (2004).





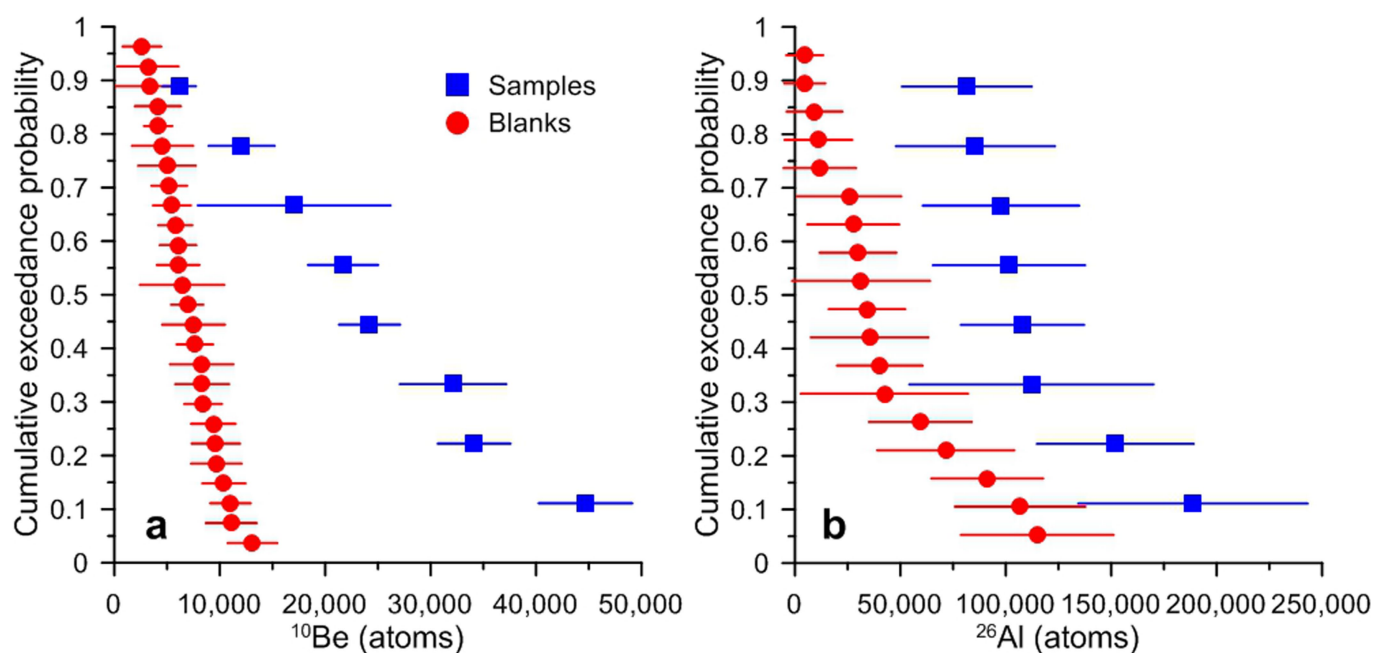
**Extended Data Fig. 1 | Modelled patterns of erosion.** a–d, Simulated erosion potential under the Antarctic Ice Sheet, calculated from modelled driving stress and basal velocity fields for several uniform (atmosphere and ocean) warming scenarios of  $4^{\circ}\text{C}$  (a),  $8^{\circ}\text{C}$  (b),  $12^{\circ}\text{C}$  (c) and  $15^{\circ}\text{C}$  (d)<sup>55</sup>.

The location of the AND-1B core is shown by the yellow dot. We note that erosive zones tend to extend towards the continental interior with warming.  $dT$ , temperature anomaly from present;  $dV$ , ice-volume anomaly from present, in sea-level equivalent (s.l.e.).



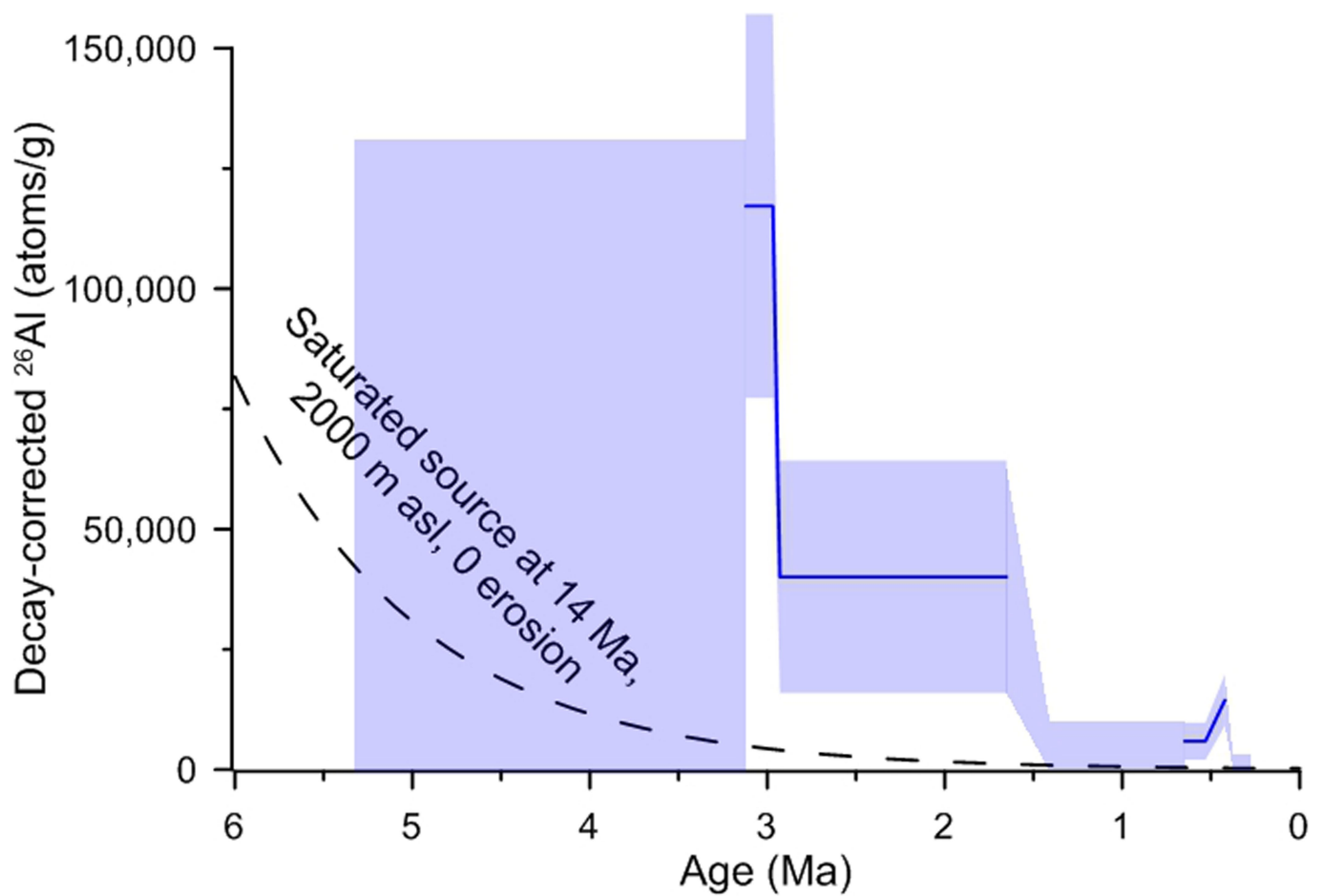
**Extended Data Fig. 2 | Glacial isostatic adjustment following ice retreat.** **a–d**, Antarctic land above sea level (yellow) 0 kyr (**a**), 5 kyr (**b**), 10 kyr (**c**), and 15 kyr (**d**) after a near-instantaneous (1-kyr) collapse of all marine-based ice-sheet sectors, in two different models of mantle viscosity<sup>26</sup>.

Model 1 is from ref. <sup>56</sup>, and model 2 (our model) has the following parameters: lithosphere thickness, 96 km; upper-mantle viscosity,  $5 \times 10^{20} \text{ Pa s}^{-1}$ ; and lower-mantle viscosity,  $10^{22} \text{ Pa s}^{-1}$ . The location of the AND-1B core is shown by the star.



**Extended Data Fig. 3 | Nuclide abundances in AND-1B samples versus blank populations. a, b,** Cumulative exceedance probabilities of measured (that is, not blank-corrected)  $^{10}\text{Be}$  (a) and  $^{26}\text{Al}$  (b) nuclide abundances in AND-1B samples (blue) and in all blanks run by the same operator in the same low-level fume hood (red), with  $1\sigma$  uncertainties. These plots display

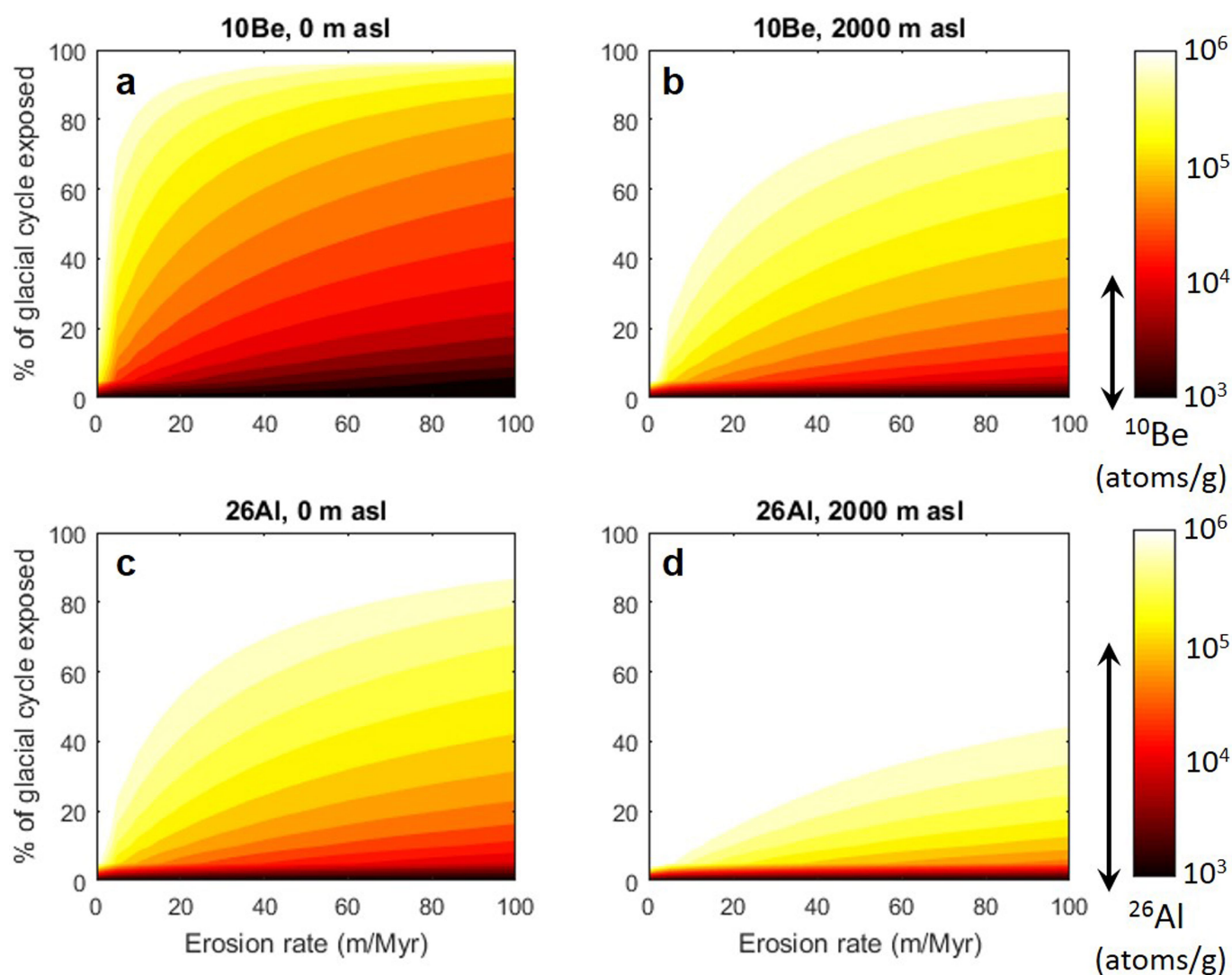
the fraction of measurements that exceed a given nuclide abundance. Note that probabilities are generally higher for the samples than the blanks; in other words, a random draw from the samples is more likely to be above a random draw from the blanks, suggesting that they are separable populations.



**Extended Data Fig. 4 | AND-1B decay-corrected  $^{26}\text{Al}$  concentrations.** Shaded intervals surrounding the blue line show  $1\sigma$  uncertainties, while shaded intervals not surrounding the blue line show the possible range of decay-corrected concentrations in samples that are below the detection limit. The dashed black line simulates the  $^{26}\text{Al}$  concentration in non-eroding material at 2,000 metres above sea level (m asl) that was

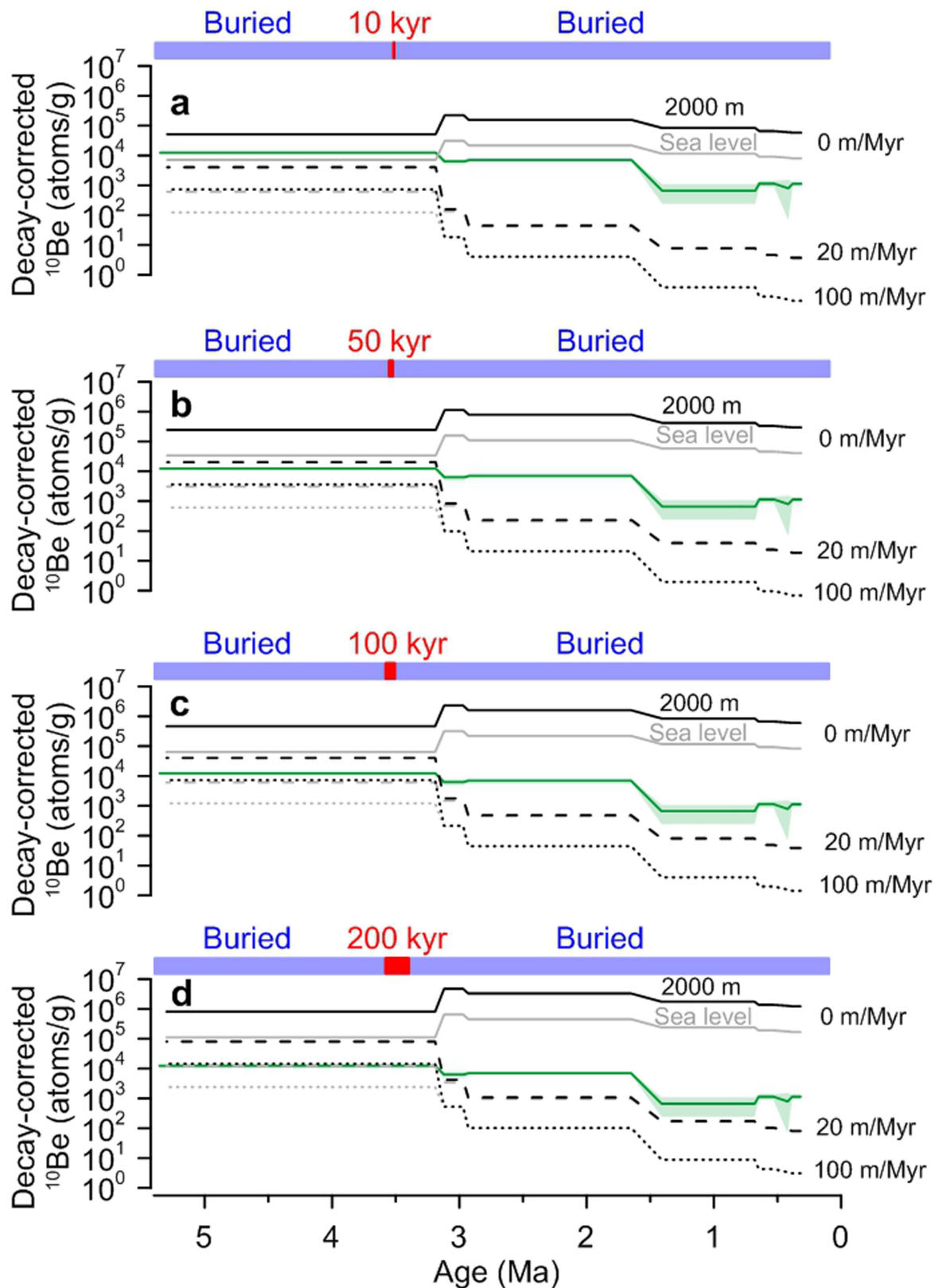
originally saturated at 14 Ma and subsequently decayed under cold-based, non-erosive ice. The fact that several AND-1B samples have higher concentrations than those in this extreme scenario (which is the most favourable to having nuclides persist to the present) suggests that the AND-1B nuclides were produced after the expansion of the EAIS in the mid-Miocene.





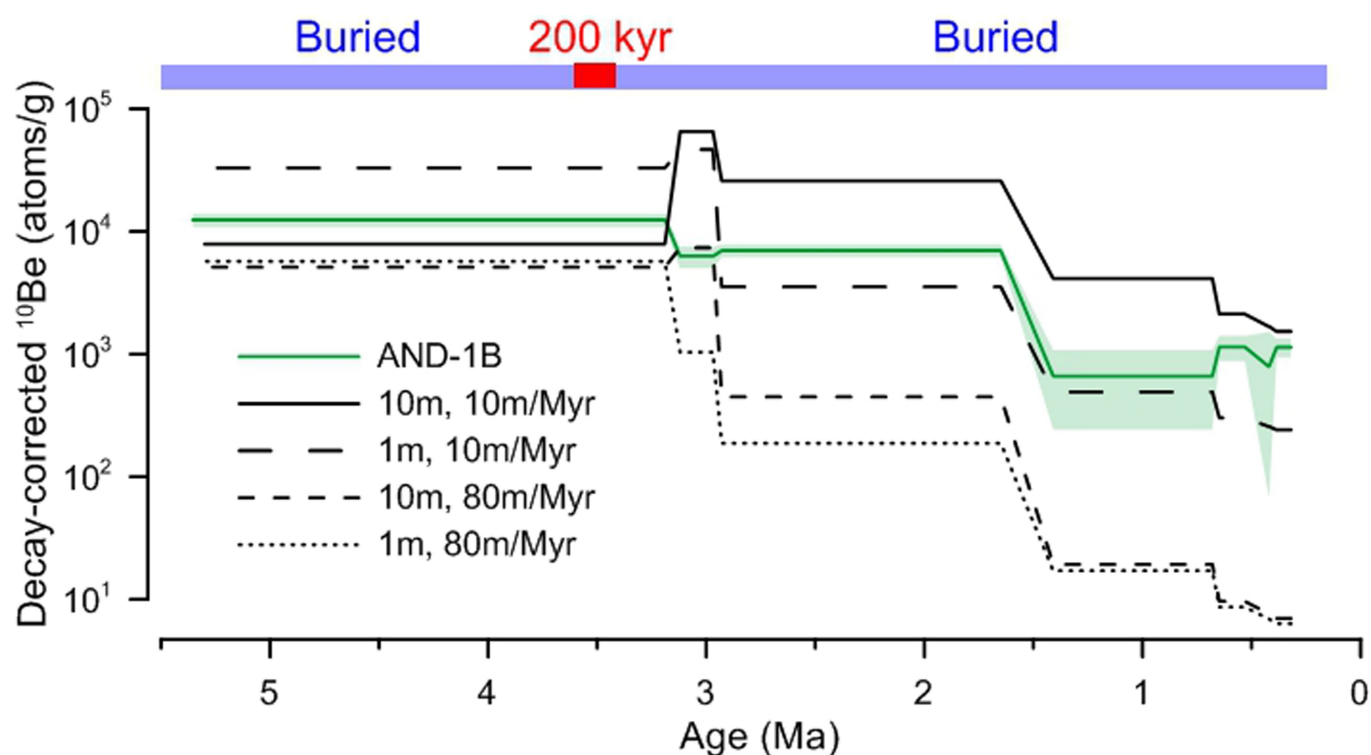
**Extended Data Fig. 5 | Modelled concentrations of cosmogenic nuclides for various durations of interglacial exposure and glacial erosion rates.** a–d, Simulated  $^{10}\text{Be}$  (a, b) and  $^{26}\text{Al}$  (c, d) concentrations in material sourced from sea level and from 2,000 m asl in Antarctica as a function of the fraction of time for which land is exposed, during 40-kyr glacial cycles. (Results are nearly identical if the cycles are instead 100-kyr long.) Erosion rates were assumed to be 0 m per Myr during ice-free conditions, on the basis of geologic evidence for negligible late Cenozoic erosion in ice-free areas of the TAMs<sup>9,10</sup>. Black arrows next to the scale bars show the range of

decay-corrected nuclide concentrations in AND-1B samples. The model was initialized with zero nuclides at 8 Ma (representative of conditions suggested by AND-1B sample H); the model also assumes instantaneous transport of eroded sediment to the ocean with no mixing, and continuous radioactive decay. Concentrations shown are the Pliocene (5 Ma to 3 Ma) average. Comparison of these simulations with AND-1B nuclide concentrations suggests that land exposure in sediment source regions was probably quite limited in duration or extent through the Plio-Pleistocene.



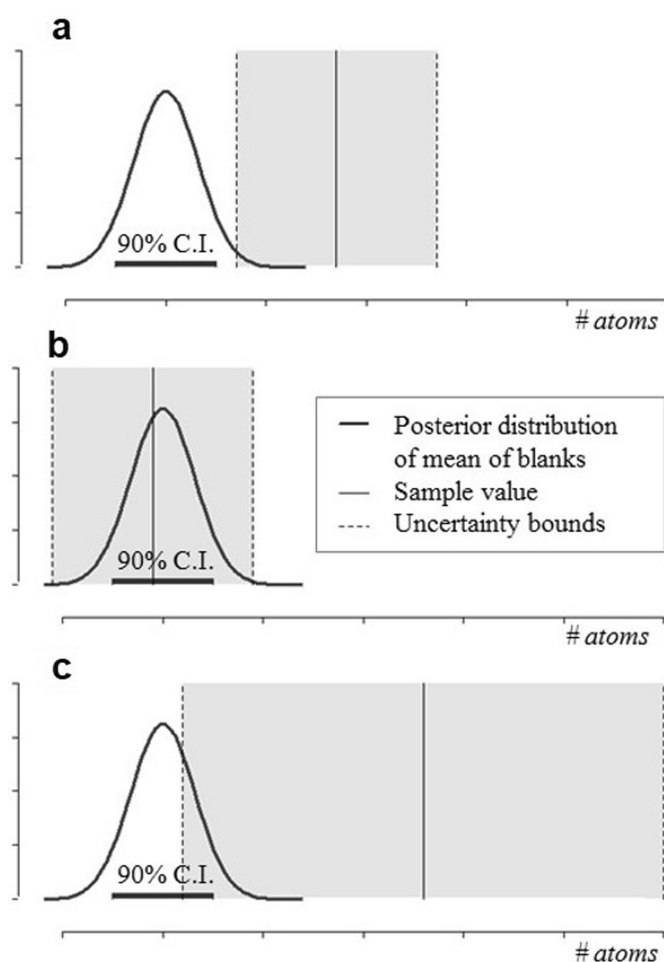
**Extended Data Fig. 6 | Modelling a mid-Pliocene exposure of bedrock.** a–d, Each panel shows actual AND-1B decay-corrected  $^{10}\text{Be}$  concentrations with  $1\sigma$  uncertainty (green), as well as simulated  $^{10}\text{Be}$  concentrations assuming a single 10-kyr (a), 50-kyr (b), 100-kyr (c) and 200-kyr (d) exposure of a bedrock column in the mid-Pliocene. The exposure event was chosen to start at 3.6 Ma and extend for up to 200 kyr in duration on the basis of the presence of a 60-m-thick diatomite unit in the AND-1B core, thought to reflect warm interglacial conditions

from 3.6 Ma to 3.4 Ma<sup>1</sup>. Simulated records are driven by production at sea level (grey) or at 2,000 m asl (black), and are subjected to continuous radioactive decay and continuous erosion at rates of 0 m per Myr (solid lines), 20 m per Myr (dashed lines), and 100 m per Myr (dotted lines). The model assumes that the sediment source was initially devoid of nuclides and that sediments are transported instantaneously to the sea floor. The synthetic time series have been binned to the same resolution as the AND-1B data.



**Extended Data Fig. 7 | Modelling a mid-Pliocene exposure event with eroded bedrock mixed through a deformable bed.** The figure shows AND-1B decay-corrected  $^{10}\text{Be}$  concentrations with  $1\sigma$  uncertainties (green). It also depicts simulated  $^{10}\text{Be}$  concentrations, assuming a single exposure event from 3.6 Ma to 3.4 Ma and routing of eroded bedrock through a well mixed deformable bed, for various bed thicknesses and erosion rates. Material eroded from the bedrock profile is instantaneously mixed throughout the deformable bed in each time step, and an equal amount of material is removed from the bed, keeping its thickness

constant. Sediment mixing in the deformable bed dilutes the surface  $^{10}\text{Be}$  signal of the exposure event but extends its longevity through time in comparison with the bedrock simulations shown in Extended Data Fig. 6. Simulated records are driven by production at sea level, and subjected to continuous radioactive decay and continuous erosion. The model assumes that the bedrock and deformable bed were initially devoid of nuclides and that sediments eroded from the deformable bed are transported instantaneously to the sea floor. The synthetic time series have been binned to the same resolution as the AND-1B data.



**Extended Data Fig. 8 | Conceptual diagram showing the outcomes of Bayesian one-group *t*-tests and their interpretation.** **a**, Nuclides are credibly present above background: that is, the sample value is greater than the mean of the blanks (defined at the mode of the posterior distribution), and the region of uncertainty surrounding the sample value fully excludes the 90% credible interval (C.I.) on the posterior distribution of the mean of the blanks. The grey shaded regions give the uncertainty range in the sample nuclide concentration. **b**, Nuclides are not credibly present above background: the sample value is less than or equal to the blank mean. **c**, Nuclides are not credibly present above background: although the sample value is greater than the blank mean, the region of uncertainty surrounding the sample value does not fully exclude the 90% C.I.



Extended Data Table 1 | Comparison of nuclide abundances in sample populations and procedural blank populations

Test	<sup>10</sup> Be			<sup>26</sup> Al		
	Blanks (10 <sup>3</sup> atoms)	Samples (10 <sup>3</sup> atoms)	Two-sample, one- sided t-test ( $\alpha=0.05$ )	Blanks (10 <sup>4</sup> atoms)	Samples (10 <sup>4</sup> atoms)	Two-sample, one- sided t-test ( $\alpha=0.05$ )
<b>Frequentist<sup>†</sup></b>	7.05±0.54 n=26	24.01±4.45 n=8	Frequentist: p = 0.0033*  Rejects H <sub>0</sub> (that mean of the samples is less than or equal to the mean of the blanks)	4.18±0.80 n=18	11.59±1.29 n=8	Frequentist: p < 0.0001*  Rejects H <sub>0</sub> (that mean of the samples is less than or equal to the mean of the blanks)
<b>Bayesian<sup>‡</sup></b>	7.02±0.58 n=26	23.91±5.92 n=8	Bayesian: 99.4% probability that sample mean is greater than the blank mean	3.98±0.89 n=18	11.40±1.68 n=8	Bayesian: 99.9% probability that sample mean is greater than the blank mean

\*The mean sample nuclide concentration is higher than the mean blank nuclide concentration at the 95% confidence level.

<sup>†</sup>Arithmetic mean and standard error of the mean of samples and of all blanks associated with the fume hood in which samples were processed by the same operator.

<sup>‡</sup>Bayesian estimation of mean and standard error of the mean of samples and of all blanks associated with the fume hood in which samples were processed by the same operator.

We note that the mean and standard error values estimated using frequentist and Bayesian methods differ slightly, given that the former is derived arithmetically and the latter is derived through MCMC sampling.

Extended Data Table 2 | Comparison of nuclide abundances in individual samples and procedural blanks

Sample	<sup>10</sup> Be				
	Measured (10 <sup>3</sup> atoms)	Corrected with blanks using Freq mean/SEM (10 <sup>3</sup> atoms) <sup>†</sup>	Frequentist one- sample t-test ( $\alpha=0.10$ ) <sup>‡</sup>	Corrected with blanks using Bayes mean/SEM (10 <sup>3</sup> atoms) <sup>§</sup>	Bayesian one-sample t- test 90% Credible Interval / ROU = 1 $\sigma$ <sup>  </sup>
A	24.21±2.91	17.16±2.96	p < 0.10*	17.19±2.97	0/0/Y
B	17.06±9.13	10.01±9.15	p < 0.10*	10.04±9.15	0/5.7/Y
C	21.68±3.33	14.63±3.38	p < 0.10*	14.66±3.38	0/0/Y
D	12.07±3.14	5.02±3.18	p < 0.10*	5.05±3.19	0/0.1/Y
E	44.67±4.45	37.62±4.49	p < 0.10*	37.65±4.49	0/0/Y
F	32.12±5.05	25.07±5.08	p < 0.10*	25.10±5.08	0/0/Y
G	34.11±3.46	27.06±3.51	p < 0.10*	27.09±3.51	0/0/Y
H	6.167±1.58	- 0.88±1.67	p = 0.943	- 0.85±1.68	92.9/89.9/N

Sample	<sup>26</sup> Al				
	Measured (10 <sup>4</sup> atoms)	Corrected with blanks using Freq mean/SEM (10 <sup>4</sup> atoms) <sup>†</sup>	Frequentist one- sample t-test ( $\alpha=0.10$ ) <sup>‡</sup>	Corrected with blanks using Bayes mean/SEM (10 <sup>4</sup> atoms) <sup>§</sup>	Bayesian one-sample t- test 90% Credible Interval / ROU = 1 $\sigma$ <sup>  </sup>
A	8.56±3.77	4.39±3.86	p < 0.10*	4.59±3.88	0/17.5/N
B	18.87±5.44	14.69±5.50	p < 0.10*	14.89±5.51	0/0/Y
C	9.77±3.71	5.59±3.79	p < 0.10*	5.80±3.81	0/1.3/Y
D	8.15±3.08	3.97±3.18	p < 0.10*	4.17±3.21	0/10.8/N
E	10.15±3.61	5.97±3.70	p < 0.10*	6.18±3.72	0/0.4/Y
F	15.19±3.72	11.02±3.80	p < 0.10*	11.22±3.82	0/0.0/Y
G	11.21±5.79	7.03±5.85	p < 0.10*	7.23±5.86	0/5.4/Y
H	10.79±2.92	6.61±3.02	p < 0.10*	6.81±3.05	0/0.0/Y

\*The sample nuclide concentration is higher than the mean blank nuclide concentration at the 90% confidence level.

<sup>†</sup>Arithmetic average of blanks subtracted from measured atoms in each sample, with uncertainties added in quadrature. In calculating the blank average, a value of one-half the lowest detected blank value was substituted for the first two of the <sup>26</sup>Al blanks, which exhibited zero-count results.

<sup>‡</sup>The test result is significant at the specified  $\alpha$  level where indicated by asterisks. The test compared non-corrected samples individually with the mean of the blanks.

<sup>§</sup>The Bayesian-estimated blank mean was subtracted from the number of measured atoms in each sample, with uncertainties added in quadrature, using the Bayesian estimated standard error of the mean (SEM) of the blanks.

<sup>||</sup>Values formatted as 'a/b/c' represent the following statistics: a, the percentage of the posterior distribution of the mean of the blanks that is above the sample value; b, the percentage of the posterior distribution of the mean of the blanks that is within the region of uncertainty (ROU) around the sample value (where the ROU is defined as  $\pm 1\sigma$ ); c, Y = yes and N = no in answer to the question

"Constituent credibly present in sample?" (that is, is the 90% credibility interval of the posterior distribution of the mean of the blanks less than and fully excluding the ROU around the sample value?).

# Rapid recovery of life at ground zero of the end-Cretaceous mass extinction

Christopher M. Lowery<sup>1\*</sup>, Timothy J. Bralower<sup>2</sup>, Jeremy D. Owens<sup>3</sup>, Francisco J. Rodríguez-Tovar<sup>4</sup>, Heather Jones<sup>2</sup>, Jan Smit<sup>5</sup>, Michael T. Whalen<sup>6</sup>, Philippe Claeys<sup>7</sup>, Kenneth Farley<sup>8</sup>, Sean P. S. Gulick<sup>1</sup>, Joanna V. Morgan<sup>9</sup>, Sophie Green<sup>10</sup>, Elise Chenot<sup>11</sup>, Gail L. Christeson<sup>1</sup>, Charles S. Cockell<sup>12</sup>, Marco J. L. Coolen<sup>13</sup>, Ludovic Ferrière<sup>14</sup>, Catalina Gebhardt<sup>15</sup>, Kazuhisa Goto<sup>16</sup>, David A. Kring<sup>17</sup>, Johanna Lofi<sup>18</sup>, Rubén Ocampo-Torres<sup>19</sup>, Ligia Perez-Cruz<sup>20</sup>, Annemarie E. Pickersgill<sup>21,22</sup>, Michael H. Poelchau<sup>23</sup>, Auriol S. P. Rae<sup>9</sup>, Cornelia Rasmussen<sup>1</sup>, Mario Rebolledo-Vieyra<sup>24</sup>, Ulrich Riller<sup>25</sup>, Honami Sato<sup>26</sup>, Sonia M. Tikoo<sup>27</sup>, Naotaka Tomioka<sup>28</sup>, Jaime Urrutia-Fucugauchi<sup>20</sup>, Johan Vellekoop<sup>7</sup>, Axel Wittmann<sup>29</sup>, Long Xiao<sup>30</sup>, Kosei E. Yamaguchi<sup>31,32</sup> & William Zylberman<sup>33</sup>

**The Cretaceous/Palaeogene mass extinction eradicated 76% of species on Earth<sup>1,2</sup>. It was caused by the impact of an asteroid<sup>3,4</sup> on the Yucatán carbonate platform in the southern Gulf of Mexico 66 million years ago<sup>5</sup>, forming the Chicxulub impact crater<sup>6,7</sup>. After the mass extinction, the recovery of the global marine ecosystem—measured as primary productivity—was geographically heterogeneous<sup>8</sup>; export production in the Gulf of Mexico and North Atlantic–western Tethys was slower than in most other regions<sup>8–11</sup>, taking 300 thousand years (kyr) to return to levels similar to those of the Late Cretaceous period. Delayed recovery of marine productivity closer to the crater implies an impact-related environmental control, such as toxic metal poisoning<sup>12</sup>, on recovery times. If no such geographic pattern exists, the best explanation for the observed heterogeneity is a combination of ecological factors—trophic interactions<sup>13</sup>, species incumbency and competitive exclusion by opportunists<sup>14</sup>—and ‘chance’<sup>8,15,16</sup>. The question of whether the post-impact recovery of marine productivity was delayed closer to the crater has a bearing on the predictability of future patterns of recovery in anthropogenically perturbed ecosystems. If there is a relationship between the distance from the impact and the recovery of marine productivity, we would expect recovery rates to be slowest in the crater itself. Here we present a record of foraminifera, calcareous nannoplankton, trace fossils and elemental abundance data from within the Chicxulub crater, dated to approximately the first 200 kyr of the Palaeocene. We show that life reappeared in the basin just years after the impact and a high-productivity ecosystem was established within 30 kyr, which indicates that proximity to the impact did not delay recovery and that there was therefore no impact-related environmental control on recovery. Ecological processes probably controlled the recovery of productivity after the Cretaceous/Palaeogene mass extinction and are therefore likely to be important for the response of the ocean ecosystem to other rapid extinction events.**

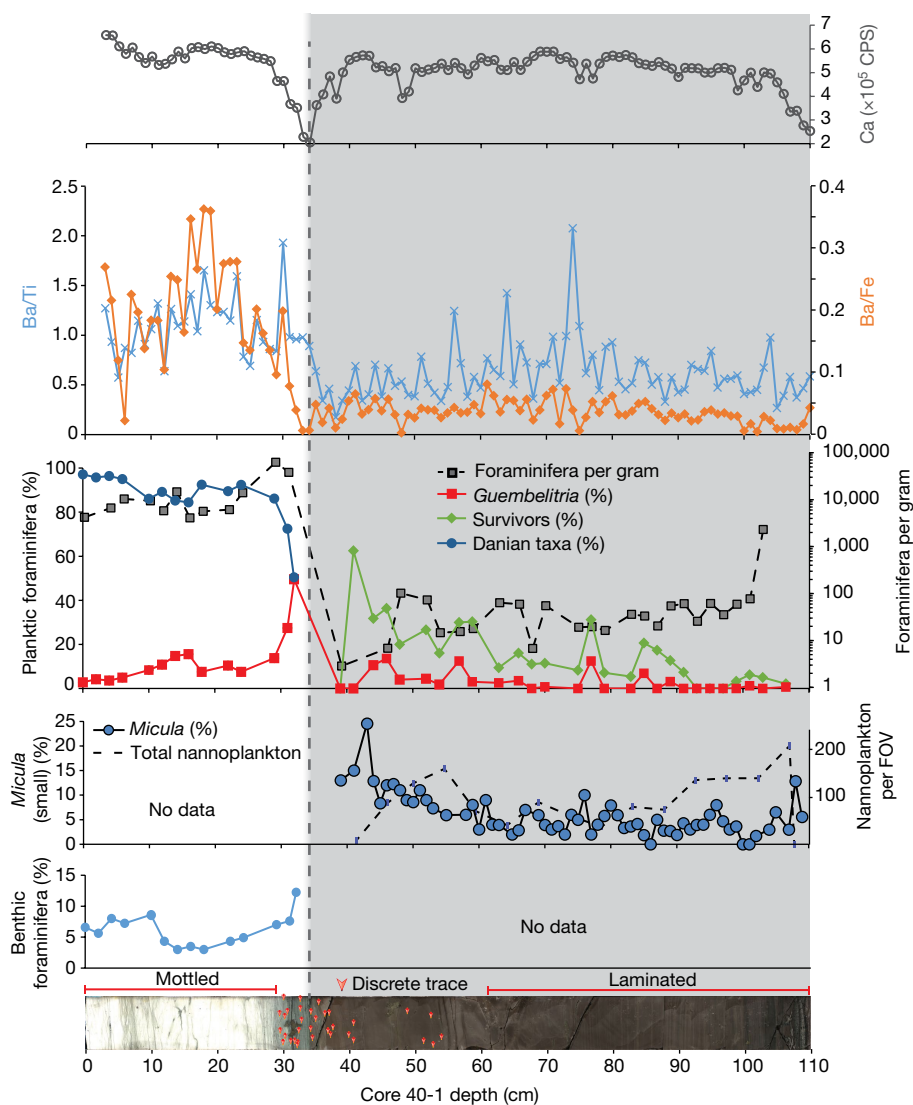
The recent joint expedition of the International Ocean Discovery Program and International Continental Drilling Program (hereafter,

Expedition 364) recovered what is, to our knowledge, the first record of the few hundred thousand years immediately after the impact within the Chicxulub crater. Site M0077, which was drilled into the peak ring of the crater<sup>7</sup> (Extended Data Fig. 1), sampled an approximately 130-m-thick, generally upward-fining suevite (that is, melt-bearing impact breccia) overlying impact melt rocks and fractured granite<sup>17</sup>. The boundary between the suevite and overlying earliest-Palaeocene pelagic limestone is in core 40-1 (Fig. 1), and comprises a 76-cm-thick upward-fining, brown, fine-grained micritic limestone that we term the ‘transitional unit’. The lower portion of the transitional unit is laminated below 54-cm core depth and contains no trace fossils (Fig. 1 and Extended Data Fig. 2). The laminations are thin, graded beds with sub-millimetre-scale cross-bedding that indicates bottom currents, and are likely due to the movement of wave energy—including tsunami and/or seiches—in the days after the impact. The fine grain size (primarily clay to silt, with some sand-sized grains concentrated in the graded beds) suggests that much of the material in the transitional unit was deposited from resuspension and settling. The transitional unit is overlain by a white pelagic limestone. The lowermost sample taken in this limestone (34 cm core depth) contains the planktic foraminifer *Parvularugoglobigerina eugubina* (which marks the base of Zone P $\alpha$ ), other foraminifer of the same genus (*P. extensa*, *P. alabamensis*) and *Guembelitra cretacea*. Because many other species that originate within Zone P $\alpha$  first appear a few centimetres higher in the section (31–32 cm), we conclude that the base of the limestone lies very near the base of this zone, 30 kyr after the impact<sup>18</sup>.

Biostratigraphy and basic assumptions about depositional and crater processes indicate that the transitional unit was deposited between several years and 30 kyr after impact (Fig. 2). To better constrain this, we use the abundance of extraterrestrial <sup>3</sup>He to determine sediment accumulation rates (see Methods). This proxy provides a firm upper limit of 8 kyr for deposition, assuming none of the <sup>3</sup>He is reworked. If even a small amount of <sup>3</sup>He is reworked (which is very likely given the prevalence of reworked microfossils and impact debris), then the transitional unit was deposited in a period of time of less than about

<sup>1</sup>Institute for Geophysics, Jackson School of Geosciences, University of Texas at Austin, Austin, TX, USA. <sup>2</sup>Department of Geosciences, Pennsylvania State University, University Park, PA, USA.

<sup>3</sup>Department of Earth, Ocean and Atmospheric Science and National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL, USA. <sup>4</sup>Departamento de Estratigrafía y Paleontología, Universidad de Granada, Granada, Spain. <sup>5</sup>Faculty of Earth and Life Sciences (FALW), Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. <sup>6</sup>Department of Geosciences, University of Alaska Fairbanks, Fairbanks, AK, USA. <sup>7</sup>Analytical, Environmental and Geo-Chemistry, Vrije Universiteit Brussel, Brussels, Belgium. <sup>8</sup>Division of Geological and Planetary Sciences, MS 170-25, California Institute of Technology, Pasadena, CA, USA. <sup>9</sup>Department of Earth Science and Engineering, Imperial College London, London, UK. <sup>10</sup>British Geological Survey, Edinburgh, UK. <sup>11</sup>Biogéosciences Laboratory, Université de Bourgogne-Franche Comté, Dijon, France. <sup>12</sup>UK Centre for Astrobiology, School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK. <sup>13</sup>School of Earth and Planetary Sciences, WA-Organic and Isotope Geochemistry Centre (WA-OIGC), Curtin University, Bentley, Western Australia, Australia. <sup>14</sup>Natural History Museum, Vienna, Austria. <sup>15</sup>Alfred Wegener Institute, Helmholtz Centre of Polar and Marine Research, Bremerhaven, Germany. <sup>16</sup>International Research Institute of Disaster Science, Tohoku University, Sendai, Japan. <sup>17</sup>Lunar and Planetary Institute, Houston, TX, USA. <sup>18</sup>Géosciences Montpellier, CNRS, Université de Montpellier, Montpellier, France. <sup>19</sup>Groupe de Physico-Chimie de l'Atmosphère, L'Institut de Chimie et Procédés pour l'Énergie, l'Environnement et la Santé (ICPEES), Université de Strasbourg, Strasbourg, France. <sup>20</sup>Instituto de Geofísica, Universidad Nacional Autónoma de México, Mexico City, Mexico. <sup>21</sup>School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK. <sup>22</sup>Argon Isotope Facility, Scottish Universities Environmental Research Centre (SUERC), East Kilbride, UK. <sup>23</sup>Department of Geology, University of Freiburg, Freiburg, Germany. <sup>24</sup>Independent consultant, Cancun, Mexico. <sup>25</sup>Institut für Geologie, Universität Hamburg, Hamburg, Germany. <sup>26</sup>Ocean Resources Research Center for Next Generation, Chiba Institute of Technology, Chiba, Japan. <sup>27</sup>Earth and Planetary Sciences, Rutgers University, New Brunswick, NJ, USA. <sup>28</sup>Kochi Institute for Core Sample Research, Japan Agency for Marine-Earth Science and Technology, Kochi, Japan. <sup>29</sup>LeRoy Eyring Center for Solid State Science, Physical Sciences, Arizona State University, Tempe, AZ, USA. <sup>30</sup>Planetary Science Institute, School of Earth Sciences, China University of Geosciences, Wuhan, China. <sup>31</sup>Department of Chemistry, Toho University, Chiba, Japan. <sup>32</sup>NASA Astrobiology Institute, Mountain View, CA, USA. <sup>33</sup>CNRS, Institut pour la Recherche et le Développement, Aix Marseille University, Marseille, France. \*e-mail: cmlowery@utexas.edu

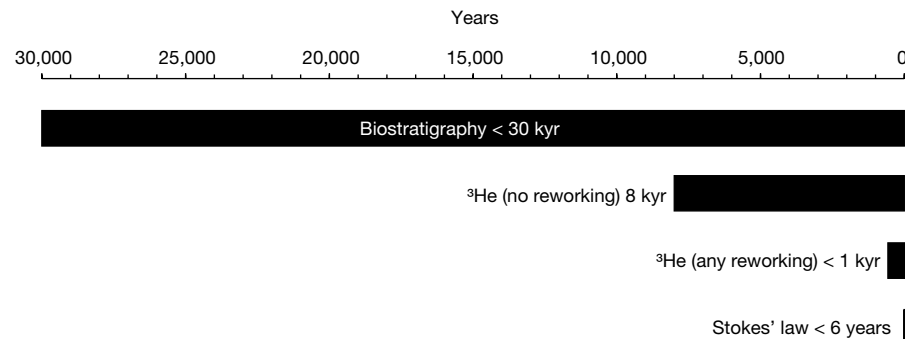


**Fig. 1 | Palaeoproductivity indicators in the earliest Palaeocene at site M0077.** The shaded area is the transitional unit and the dashed line represents the contact with the overlying pelagic limestone. Top to bottom: X-ray fluorescence-derived calcium abundance in counts per second (CPS); Ba/Ti and Ba/Fe ratios; percentage abundances of key planktic foraminiferal groups, including percentage of *Guembelitrina*, percentage of survivors (that is, Cretaceous species known to survive the impact) and percentage of Danian taxa (that is, species that evolved after the impact)

as a percentage of total foraminifera; foraminifera per gram of sediment, plotted on a logarithmic scale; percentage of *Micula* smaller than 2  $\mu\text{m}$  (against total nannoplankton) and nannoplankton abundance (total occurrences per field of view (FOV)); percentage of benthic foraminifera (against total foraminifera); and core image of 364-M0077A-40R-1 0–110 cm (616.58–617.33 m below seafloor), with discrete trace fossils highlighted by arrows (see Extended Data Fig. 2 for a larger version of this image).

1 kyr, which is below the resolution of the method. With no sediment source other than settling of material suspended by the impact and subsequent tsunami and seiches, a more realistic estimate—based on

Stokes' law—for the duration of this unit suggests about 6 years for the settling of a 2- $\mu\text{m}$  grain of carbonate (an upper limit, as most grains are much larger; see Supplementary Information for further discussion).



**Fig. 2 | Constraints on the age of the transitional unit.** Maximum durations of the transitional unit based on biostratigraphy (which suggests it was deposited in less than 30 kyr), extraterrestrial  $^3\text{He}$  (which suggests

it was deposited in approximately 8 kyr if there is no reworking, or less than 1 kyr if there is any reworking) and Stokes' law, which suggests it was deposited in less than 6 years.



The lower portion of the overlying limestone, which contains fossils that appear approximately 30 kyr after the impact, appears conformable with the transitional unit and must therefore be condensed owing to low pelagic sedimentation in the first few tens of thousands of years after the impact.

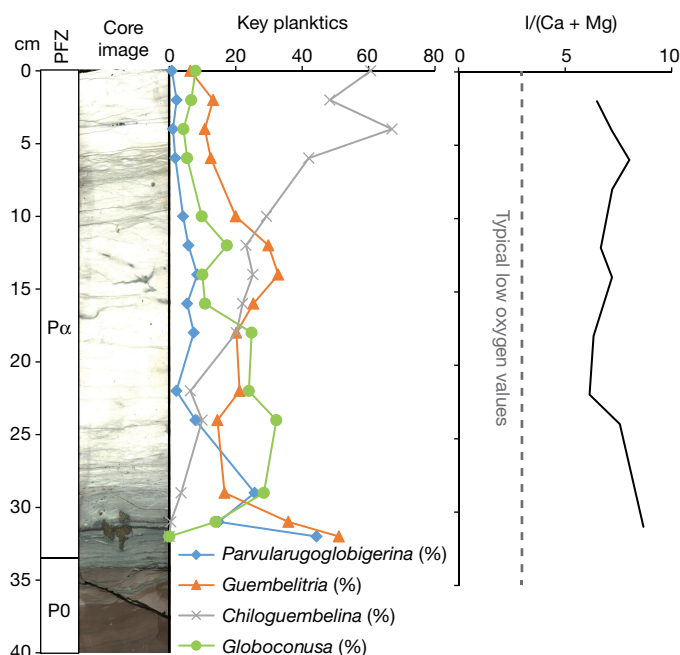
Clear, discrete trace fossils, including *Planolites* and *Chondrites*, characterize the upper 20 cm of the transitional unit (above 54 cm) (Fig. 1 and Extended Data Fig. 2), providing unequivocal evidence for benthic life in the crater within years of the impact. Flattening of the structures indicates that the traces were formed while the sediment was still soft, during or shortly after the deposition of the transitional unit. Infilling of the burrows with brown, fine-grained micrite also suggests traces were syndepositional and not derived from mixing of the Danian limestone above the transitional unit. Trace fossils produced during deposition of the limestone, as indicated by light infilling material, are distinct and occur only in the uppermost few centimetres of the transitional unit (Extended Data Fig. 2).

The transitional unit microfossils are dominated by clearly reworked Maastrichtian foraminifera and nannoplankton, known across the Gulf of Mexico and Caribbean as the Cretaceous/Palaeogene (K/Pg) boundary cocktail<sup>19</sup> (Extended Data Fig. 3 and Supplementary Table 1). Although overall foraminiferal abundance (plotted as the number of foraminifera per gram of sedimentary rock; Fig. 1) is high at the base of the unit, species known to range across the boundary ('survivor species') are rare in the lower transitional unit and become more common up-section even as total foraminifera decline (Fig. 1). Survivor species, here defined as *G. cretacea*, *Muricohedbergella monmouthensis* and *Muricohedbergella holmdelensis*<sup>20</sup>, dominate a depauperate assemblage in the upper 20 cm of the transitional unit, coinciding with the first appearance of trace fossils (Extended Data Figs. 4, 5).

The nannofossil assemblage in the transitional unit contains reworked Cretaceous specimens, including a group of clearly overgrown species (such as *Aspidolithus parvus* (also known as *Broinsonia parva*) and *Eiffellithus eximius*) that became extinct near the Campanian/Maastrichtian boundary. The remainder of the Cretaceous species, which dominate the assemblage, range to the top of or beyond the latest Maastrichtian age (Supplementary Table 2). Unusually small (less than 2 µm) and delicate specimens of *Micula* are observed throughout the transitional unit and increase in abundance up-section (Fig. 1), along with small *Retecapsa* (Extended Data Fig. 6). Taxa common at other sites of the earliest Danian stage are also present, including disaster genera (opportunistic groups that can tolerate high environmental stress) such as *Thoracosphaera* and *Braarudosphaera*. Unlike the foraminifera, there are no clear stratigraphic trends in overall nannoplankton abundance (Fig. 1).

Because survivor species lived both before and after the K/Pg mass extinction, it is impossible to determine for certain whether individual specimens in the transitional unit colonized the crater after the impact. However, the populations of foraminifera and nannoplankton are substantially different from those of the latest Cretaceous<sup>12</sup> (that is, the expected population if the whole assemblage was reworked), suggesting that these taxa were true survivors (Fig. 1 and Extended Data Fig. 6). *G. cretacea*, a common component of the survivor assemblage in the upper transitional unit, was restricted to marginal marine waters during the Maastrichtian and would not have been present at the pre-impact site, which was over 100 m deep<sup>21</sup> and over 500 km from shore<sup>22</sup>. The nannofossil assemblage in the transitional unit is considerably different from typical latest Maastrichtian assemblages, with some genera over-represented (*Watznaueria* and *Retecapsa*) and others under-represented (*Eiffellithus*, not including *E. eximius*, *Arkhangelskiella*, *Chiastozygus* and *Prediscosphaera*) (Extended Data Fig. 6). Additionally, *Micula*—a robust taxon often used as a proxy for dissolution—is not as abundant as elsewhere, indicating that these unusual abundances are not due to poor or selective preservation (Extended Data Fig. 6).

This initial appearance of life is notably fast, especially because crater-specific factors do not seem to have had a negative effect on the



**Fig. 3 | Early Danian foraminifer abundances and I/(Ca + Mg)**

**oxygenation proxy.** Left plot, Key Danian planktic foraminifera.

Normal perforate planktic foraminifera (*Eoglobigerina*, *Globanomalina*, *Parasubbotina* and *Praemurica*) are rare throughout the study interval and not plotted here; all are plotted as a percentage of total planktic foraminifera. Right plot, I/(Ca + Mg) redox proxy, indicating well-oxygenated conditions in the Chicxulub crater through this interval. PFZ, planktic foraminifer zone.

local recovery of life. A vigorous, high-temperature hydrothermal system was established within the crater and may have persisted for millions of years after the impact<sup>23</sup>, especially across the peak ring where rocks exhumed from deep in the crust were extensively fractured<sup>7</sup>. Nevertheless, the appearance of burrowing organisms within years of the impact indicates that the hydrothermal system did not adversely affect seafloor life. Impact-generated hydrothermal systems are hypothesized to be potential habitats for early life on Earth<sup>24</sup> and on other planets, particularly below the surface. However, for marine impact craters in open ocean communication, such as Chicxulub (Extended Data Fig. 1), our data indicate that locally substantial but comparatively small volumes of hydrothermal fluids were overwhelmed by the  $1.3 \times 10^4 \text{ km}^3$  of well-mixed ocean water that filled the basin.

Likewise, the open connection with the Gulf of Mexico prevented the development of anoxia in the crater. Our analyses of I/Ca ratios suggest that local dissolved oxygen was high and stable in Zone P $\alpha$  (Fig. 3). This is in contrast to the smaller (85-km wide) Eocene Chesapeake Bay impact crater, where anoxia due to restriction is attributed as the cause of delayed recovery of the benthic ecosystem on the crater floor<sup>25</sup>. This comparison suggests that the establishment of life within marine impact craters is controlled more by circulation (and thus crater geometry) than by the magnitude of the impact or global environmental effects.

The overlying pelagic limestone, which was deposited within Zone P $\alpha$  (30–200 kyr after the impact) contains abundant evidence of high productivity in a thriving ecosystem. The assemblage of planktic foraminifera in Zone P $\alpha$  is diverse and abundant (Fig. 3). Good preservation in the lowermost sample (34 cm core depth) enabled the identification of over 60 species of benthic foraminifera, and benthics make up 12% of the total foraminiferal assemblage at this level (Supplementary Table 1). This percentage of benthics<sup>26</sup> and the overall benthic assemblage<sup>27</sup> are both typical of a palaeo-water depth of about 600–700 m (around the boundary between the upper and middle bathyal zones)<sup>10,27</sup>. At the base of the white limestone, trace fossils increase in size, abundance and diversity relative to the underlying

transitional unit. The abundance and diversity of benthic organisms indicate that by about 30 kyr after the impact, seafloor conditions had returned to normal and sufficient organic matter flux existed to sustain a diverse, multilayer benthic community.

Conversely, the nannoplankton assemblage in the Danian limestone is dominated by *Braarudosphaera* and calcareous dinoflagellate cysts (for example, *Thoracosphaera*), which are common disaster taxa in the early recovery interval. Large, foraminifer-sized calcispheres appear after about 100 kyr. Calcareous phytoplankton in the earliest Danian clearly represent a low-diversity, high-productivity bloom. Genera such as *Neobiscutum* and *Prinsius*, which are common bloom taxa in the recovery interval at other Northern Hemisphere sites, do not become common until several metres higher in the section, over one million years after the impact. Organic microfossils are completely absent from the study interval, probably owing to poor preservation of organic material.

Geochemical palaeoproductivity proxies, particularly Ba/Ti and Ba/Fe ratios, also indicate high productivity in the post-impact Danian limestone (Fig. 1). Ba/Ti ratios of about 1.0 at the base of the limestone (approximately 30 kyr after the impact) and about 2.0 above that (15 cm higher, or about 100 kyr after the impact) indicate relatively high and increasing productivity in the Chicxulub basin in the earliest Danian.

The recovery of productivity in the crater is faster than that at many sites, including those in the Gulf of Mexico, some of which took 300 kyr or more to recover to a similar extent<sup>8,11</sup>. Therefore, we find that proximity to the impact was not a control on recovery in marine ecosystems. The wide range of rates of recovery in the oceans show no relationship with geographic distance to the crater and so are best explained by natural ecological interactions, such as incumbency and competitive exclusion, between organisms within recovery ecosystems<sup>8,14</sup>. These trends can be used to understand the rates of recovery after other major extinction events and to predict the long-term recovery of modern ecosystems affected by pollution and climate change.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0163-6>.

Received: 27 October 2017; Accepted: 3 April 2018;

Published online: 30 May 2018

- Jablonski, D. in *Extinction Rates* (eds Lawton, J. H. & May, R. M.) 25–44 (Oxford Univ. Press, Oxford, 1995).
- Schulte, P. et al. The Chicxulub asteroid impact and mass extinction at the Cretaceous–Paleogene boundary. *Science* **327**, 1214–1218 (2010).
- Alvarez, L. W., Alvarez, W., Asaro, F. & Michel, H. V. Extraterrestrial cause for the Cretaceous–Tertiary extinction. *Science* **208**, 1095–1108 (1980).
- Smit, J. & Hertogen, J. An extraterrestrial event at the Cretaceous–Tertiary boundary. *Nature* **285**, 198–200 (1980).
- Renne, P. R. et al. Time scales of critical events around the Cretaceous–Paleogene boundary. *Science* **339**, 684–687 (2013).
- Hildebrand, A. R. et al. Chicxulub crater: a possible Cretaceous/Tertiary boundary impact crater in the Yucatán Peninsula, Mexico. *Geology* **19**, 867–871 (1991).
- Morgan, J. V. et al. The formation of peak rings in large impact craters. *Science* **354**, 878–882 (2016).
- Hull, P. M. & Norris, R. D. Diverse patterns of ocean export productivity change across the Cretaceous–Paleogene boundary: new insights from biogenic barium. *Paleoceanography* **26**, PA3205 (2011).
- Alegret, L. & Thomas, E. Cretaceous/Paleogene boundary bathyal paleo-environments in the central North Pacific (DSDP site 465), the northwestern Atlantic (ODP site 1049), the Gulf of Mexico, and the Tethys: the benthic foraminiferal record. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **224**, 53–82 (2005).
- Alegret, L., Molina, E. & Thomas, E. Benthic foraminifera at the Cretaceous–Tertiary boundary around the Gulf of Mexico. *Geology* **29**, 891–894 (2001).
- Alegret, L., Arenillas, I., Arz, J. A. & Molina, E. Foraminiferal event-stratigraphy across the Cretaceous/Paleogene boundary. *Neues Jahrb. Geol. Paläontol. Abh.* **231**, 25–50 (2004).

- Jiang, S., Bralower, T. J., Patzkowsky, M. E., Kump, L. R. & Schueth, J. D. Geographic controls on nannoplankton extinction across the Cretaceous/Paleogene boundary. *Nat. Geosci.* **3**, 280–285 (2010).
- Solé, R. V., Montoya, J. M. & Erwin, D. H. Recovery after mass extinction: evolutionary assembly in large-scale biosphere dynamics. *Philos. Trans. R. Soc. Lond. B* **357**, 697–707 (2002).
- Schueth, J. D., Bralower, T. J., Jiang, S. & Patzkowsky, M. E. The role of regional survivor incumbency in the evolutionary recovery of calcareous nannoplankton from the Cretaceous/Paleogene (K/Pg) mass extinction. *Paleobiology* **41**, 661–679 (2015).
- Hull, P. M., Norris, R. D., Bralower, T. J. & Schueth, J. D. A role for chance in marine recovery from the end-Cretaceous extinction. *Nat. Geosci.* **4**, 856–860 (2011).
- Yedid, G., Ofria, C. A. & Lenski, R. E. Selective press extinctions, but not random pulse extinctions, cause delayed ecological recovery in communities of digital organisms. *Am. Nat.* **173**, E139–E154 (2009).
- Gulick, S., Morgan, J., Mellett, C. L. & the Expedition 364 Scientists. *Expedition 364 Preliminary Report: Chicxulub: Drilling the K-Pg Impact Crater* (International Ocean Discovery Program, College Station, TX, 2017).
- Wade, B. S., Pearson, P. N., Berggren, W. A. & Pälike, H. Review and revision of Cenozoic tropical planktonic foraminiferal biostratigraphy and calibration to the geomagnetic polarity and astronomical time scale. *Earth Sci. Rev.* **104**, 111–142 (2011).
- Bralower, T. J., Paull, C. K. & Leckie, R. M. The Cretaceous–Tertiary boundary cocktail: Chicxulub impact triggers margin collapse and extensive sediment gravity flows. *Geology* **26**, 331–334 (1998).
- Olsson, D. K., Hemleben, C., Berggren, W. A. & Huber, B. T. *Atlas of Paleocene Planktonic Foraminifera* (Smithsonian Institution, Washington, 1999).
- Gulick, S. P. S. et al. Importance of pre-impact crustal structure for the asymmetry of the Chicxulub impact crater. *Nat. Geosci.* **1**, 131–135 (2008).
- Sohl, N. F., Martínez, E. R., Salmerón-Ureña, P. & Soto-Jaramillo, F. in *Geology of North America, Volume J: Gulf of Mexico Basin* (ed. Salvador, A.) 205–244 (Geological Society of America, Boulder, 1991).
- Abramov, O. & Kring, D. A. Numerical modeling of impact-induced hydrothermal activity at the Chicxulub crater. *Meteorit. Planet. Sci.* **42**, 93–112 (2007).
- Cockell, C. S. The origin and emergence of life under impact bombardment. *Phil. Trans. R. Soc. Lond. B* **361**, 1845–1856 (2006).
- Poag, C. W. in *The ICDP-USGS Deep Drilling Project in the Chesapeake Bay Impact Structure: Results from the Eyreville Core Holes* (The Geological Society of America Special Paper 458) (eds Gohn, G. S. et al.) 747–773 (Geological Society of America, Boulder, 2009).
- Leckie, R. M. & Olson, H. C. In *Micropaleontologic Proxies for Sea-level Change and Stratigraphic Discontinuities* (SEPM Special Publication 75) (eds Olson, H. C. & Leckie, R. M.) 5–19 (Society for Sedimentary Geology, Tulsa, 2003).
- Alegret, L. & Thomas, E. Upper Cretaceous and lower Paleocene benthic foraminifera from northeastern Mexico. *Micropaleontology* **47**, 269–316 (2001).

**Acknowledgements** This research used samples and data provided by the International Ocean Discovery Program (IODP). IODP Expedition 364 was jointly funded by the European Consortium for Ocean Research Drilling (ECORD) and International Continental Drilling Program (ICDP), with contributions and logistical support from the Yucatán State Government and Universidad Nacional Autónoma de México (UNAM). We thank T. Cayton for assistance with crushing and washing samples; S. Dameron, R. Moura de Mello and M. Leckie for helpful discussions on benthic foraminifer taxonomy; J. Maner for assistance with the UT ESEM laboratory and R. Martindale for assistance with petrographic microscope imaging. We are particularly grateful for assistance of the staff of the IODP Core Repository in Bremen, Germany for their assistance taking these samples and running shipboard analyses. The authors acknowledge Post-Expedition Awards from the US Science Support Program for C.M.L. and T.J.B., NSF OCE 1737351, and NASA NNX16AJ60G. Funding for F.J.R.-T. was provided by Project CGL2015-66835-P (Secretaría de Estado de I+D+I, Spain), and Scientific Excellence Unit UCE-2016-05 (Universidad de Granada).

**Reviewer information** Nature thanks B. Huber and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** All authors participated in sampling and data collection offshore and/or onshore during IODP–ICDP Expedition 364. C.M.L., T.J.B., F.J.R.-T., H.J. and J.S. collected and analysed microfossil data, M.T.W. provided detailed sedimentology, and J.D.O., P.C. and K.F. collected trace element, X-ray fluorescence and He isotope data, respectively. All authors contributed to writing and/or editing of the manuscript.

**Competing interests** The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0163-6>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0163-6>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to C.M.L.  
**Publisher's note**: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

Sample size was determined according to standard community practice (collecting approximately 300 specimens per sample, when possible). No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

The IODP–ICDP Expedition 364 drilled the peak ring of the Chicxulub crater in the spring of 2016 (Extended Data Fig. 1). Samples were taken at the Bremen IODP core repository during the Expedition 364 sampling party. Core depth in centimetres—with zero at the top of the section (616.24 m below sea floor)—are reported throughout. Core material was indurated, and ~0.5-cm quarter-rounds were cut out with a rock saw. Owing to the need to reserve core material for rare earth element geochemistry (data not shown), the lowermost ~1.5 cm of the Danian limestone was not sampled. Individual samples were subdivided for foraminifera, calcareous nannoplankton and discrete geochemical analyses.

Forty-three samples were examined for planktic and benthic foraminifera from core 40 from 0–110 cm depth. Samples were weighed, crushed with a mortar and pestle, soaked overnight (or longer) in a 10% solution of hydrogen peroxide buffered with borax and washed over a 43- $\mu$ m sieve to ensure capture of small Danian taxa. The sieve was soaked in methylene blue dye between samples to identify contaminated specimens. Samples were then dried in an oven, split to obtain a manageable volume of material, and examined for foraminifera, calcispheres, and other sand-sized particles. In the Danian limestone, at least 300 specimens were counted to establish a statistically robust population<sup>28</sup> and the rest of the residue was then examined for biostratigraphically important taxa. Low abundances in the transitional unit precluded 300-specimen counts. However, we demonstrate that our values are sufficient to reject the null hypothesis (that the observed enrichments in survivor taxa are the result of random noise) with binomial confidence limits. This calculation traditionally provides the basis for the 300-specimen ‘rule’: counting 300 specimens provides statistical confidence at a 95% confidence interval that a species that makes up 1% of the population is represented in the count<sup>28</sup>. As we show, fewer specimens are sufficient to demonstrate the presence of a survivor population in our samples. Binomial confidence limits for samples with fewer than 300 specimens are reported in Supplementary Table 1. Additionally, a single unusually well-preserved sample at the base of the post-impact limestone was examined for rare benthic species to determine the true diversity of benthic foraminifera at the base of the unit (Supplementary Table 1). Planktic foraminifer biozonation follows the P zones of Berggren and Pearson<sup>29</sup> as modified by Wade et al.<sup>18</sup>.

Ninety-seven samples were examined for nannofossils. Samples were disaggregated in water, and smear slides were made from the supernatant. Slides were observed in a transmitted light microscope at 1,600 $\times$  until at least 100 specimens were observed (Supplementary Table 2). Standard taxonomy was applied (<http://www.mikrotax.org/Nannotax3/index.php?dir=Coccolithophores>). The abundance of taxa at site M0077 was compared to a previous compilation of global K/Pg nannoplankton<sup>12</sup>.

Ichnological analysis was conducted from 0–110 cm. Ichnological observations were conducted on core material and a detailed and continuous analysis of digital images. To improve visibility of ichnological features, images were treated by a digital image methodology, based on the modification of image adjustments as levels, brightness and vibrance<sup>30,31</sup>. Ichnotaxonomical classification of trace fossils was based on the overall shape and the presence of diagnostic criteria such as size and presence of branches<sup>32</sup>. Special attention was given to the infilling material of biogenic structures.

The measurement of I/(Ca + Mg) was carried out using a procedure similar to a previously described method<sup>33</sup>. For each sample and geostandard, approximately 3–4 mg of carbonate powder was weighed out, dissolved in ~0.45 M nitric solution and then diluted using 0.1 M nitric acid and 0.5% TMAH solution. All reported measurements are from samples that had a matrix of  $50 \pm 5$  p.p.m. calcium solution to ensure the most precise iodine measurement. Dissolved samples had TMAH solution added within an hour to avoid any possible loss of volatilized iodine<sup>33</sup>. Samples were measured using an Agilent inductively coupled plasma mass spectrometer 7500 cs housed within the geochemistry group of the National High Magnetic Field Laboratory at Florida State University. A previously reported known sample, Key Largo (KL 1-1) was used to ensure reliable reproducibility. Our value of 5.51  $\mu$ mol/mol was within error of the reported value of 5.55  $\mu$ mol/mol. A previous study<sup>34</sup> found that generally low oxygen conditions correspond to ~2.6  $\mu$ mol/mol for I/(Ca + Mg). Values are reported in Supplementary Table 3.

Section 1 of core 40 was scanned with an AVAATECH XRF Core Scanner II at MARUM (Bremen, Germany) during the onshore phase of Expedition 364 (Fig. 1). The split core was covered with a 4- $\mu$ m-thick SPEX CertiPrep Ultralene foil to avoid contamination. X-ray fluorescence data were acquired with a Canberra X-PIPS silicon drift detector with 1550 eV resolution, a Canberra DAS 1000 dig-

ital spectrum analyser and an Oxford Instruments 50 W XTF011 X-ray tube with rhodium target material. X-ray spectra were processed with WIN AXIL software from Canberra Erisys at a resolution of 12 mm and a step of 10 mm. Scans were conducted at different voltages to determine a range of element concentrations: 50 kV, with a beam current of 1 mA (Ba and Sr; average dead time of 5%), and 10 kV with a beam current of 0.15 mA (Al, Si, K, Ca, Ti, Fe, Mn and S; average dead time of 11%). For each scan, sampling time was 20 s per spot.

<sup>3</sup>He is delivered to the Earth's surface by cosmic dust grains and over short time spans (about one million years) can be used as a constant flux proxy<sup>35</sup>. Previous work has shown that the K/Pg impactor was not associated with enhanced <sup>3</sup>He flux, and the mean extraterrestrial <sup>3</sup>He flux from cosmic dust accretion at the end of the Cretaceous ( $106 \times 10^{-15}$  cc (standard temperature and pressure) per g per cm<sup>2</sup> per kyr) was used to estimate the duration over which the K/Pg boundary clay was deposited at Gubbio and El Kef<sup>36</sup>. We use a similar approach here to establish the sedimentation rate of the transitional unit, which we use to develop an age model.

Helium isotope ratios and concentrations were measured on ~1-g aliquots of sediment following standard analytical procedures<sup>31</sup>. Extraterrestrial <sup>3</sup>He concentrations were computed from measured He isotopic compositions using an isotopic deconvolution model<sup>36</sup>. Results are shown in Extended Data Table 1. <sup>3</sup>He concentrations and <sup>3</sup>He/<sup>4</sup>He ratios are generally low compared to typical marine sediments of similar age<sup>37,38</sup>. Nevertheless, with the exception of the lowest sample in the transitional unit (106.5 cm), the fraction of <sup>3</sup>He attributable to an extraterrestrial source is high, ranging from ~0.70 to 0.96. The deepest sample has a similar <sup>3</sup>He concentration to other samples in the transitional unit, but ~5 times more <sup>4</sup>He. This elevated <sup>4</sup>He probably arises from a higher concentration of terrigenous <sup>4</sup>He-bearing material deposited rapidly after the impact.

We see no evidence for extraterrestrial He carried in impactor fragments, such as highly elevated and/or highly variable <sup>3</sup>He and <sup>3</sup>He/<sup>4</sup>He ratios. The absence of such a signal is consistent with either (a) the absence of impactor fragments in the material analysed or (b) the loss of extraterrestrial <sup>3</sup>He from the impactor via heating, vaporization or fusion. Note that, unlike many tracers of the impactor (such as Ir), deposition of fused or vaporized impactor will leave no trace in the sedimentary record because once He is lost into the atmosphere, it can no longer be retained in sediments.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

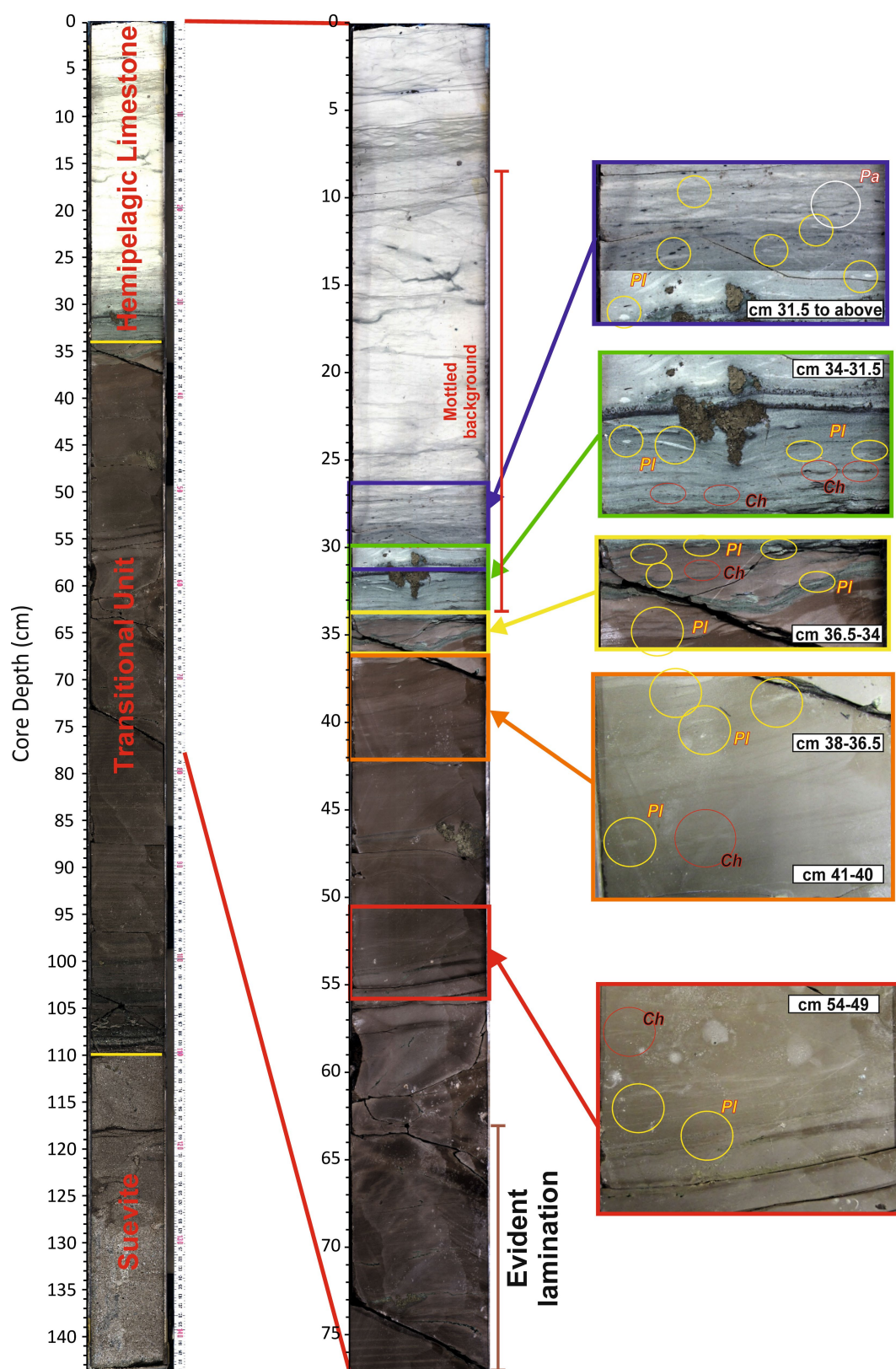
**Data availability.** X-ray fluorescence data have previously been published<sup>39</sup> and are available online (<https://doi.org/10.14379/iodp.proc.364.2017>). All other data supporting the findings of this study are available within the paper and its Supplementary Information.

28. Buzas, M. A. Another look at confidence limits for species proportions. *J. Paleontol.* **64**, 842–843 (1990).
29. Berggren, W. A. & Pearson, P. N. A revised tropical and subtropical Paleogene planktonic foraminiferal zonation. *J. Foraminiferal Res.* **35**, 279–298 (2005).
30. Dorador, J. & Rodríguez-Tovar, F. Digital image treatment applied to ichnological analysis of marine core sediments. *Facies* **60**, 39–44 (2014).
31. Dorador, J. & Rodríguez-Tovar, F. J. Stratigraphic variation in ichnofabrics at the “Shackleton Site” (IODP Site U1385) on the Iberian Margin: paleoenvironmental implications. *Mar. Geol.* **377**, 118–126 (2016).
32. Knaust, D. *Atlas of Trace Fossils in Well Core: Appearance, Taxonomy and Interpretation* (Springer, New York, 2017).
33. Lu, Z., Jenkyns, H. C. & Rickaby, R. E. M. Iodine to calcium ratios in marine carbonate as a paleo-redox proxy during oceanic anoxic events. *Geology* **38**, 1107–1110 (2010).
34. Hardisty, D. S. et al. Perspectives on Proterozoic surface ocean redox from iodine contents in ancient and recent carbonate. *Earth Planet. Sci. Lett.* **463**, 159–170 (2017).
35. Farley, K. A. & Eltgroth, S. F. An alternative age model for the Paleocene–Eocene thermal maximum using extraterrestrial <sup>3</sup>He. *Earth Planet. Sci. Lett.* **208**, 135–148 (2003).
36. Patterson, D. B. & Farley, K. A. Extraterrestrial <sup>3</sup>He in seafloor sediments: evidence for correlated 100 kyr periodicity in the accretion rate of interplanetary dust, orbital parameters, and Quaternary climate. *Geochim. Cosmochim. Acta* **62**, 3669–3682 (1998).
37. Mukhopadhyay, S., Farley, K. A. & Montanari, A. A 35 Myr record of helium in pelagic limestones from Italy: implications for interplanetary dust accretion from the early Maastrichtian to the middle Eocene. *Geochim. Cosmochim. Acta* **65**, 653–669 (2001).
38. Mukhopadhyay, S., Farley, K. A. & Montanari, A. A short duration of the Cretaceous–Tertiary boundary event: evidence from extraterrestrial helium-3. *Science* **291**, 1952–1955 (2001).
39. Morgan, J., Gulick, S., Mellet, C. L., Green, S. L. & Expedition 364 Scientists. *Chicxulub: Drilling the K-Pg Impact Crater. Proceedings of the International Ocean Discovery Program 364* (International Ocean Discovery Program, College Station, 2017).





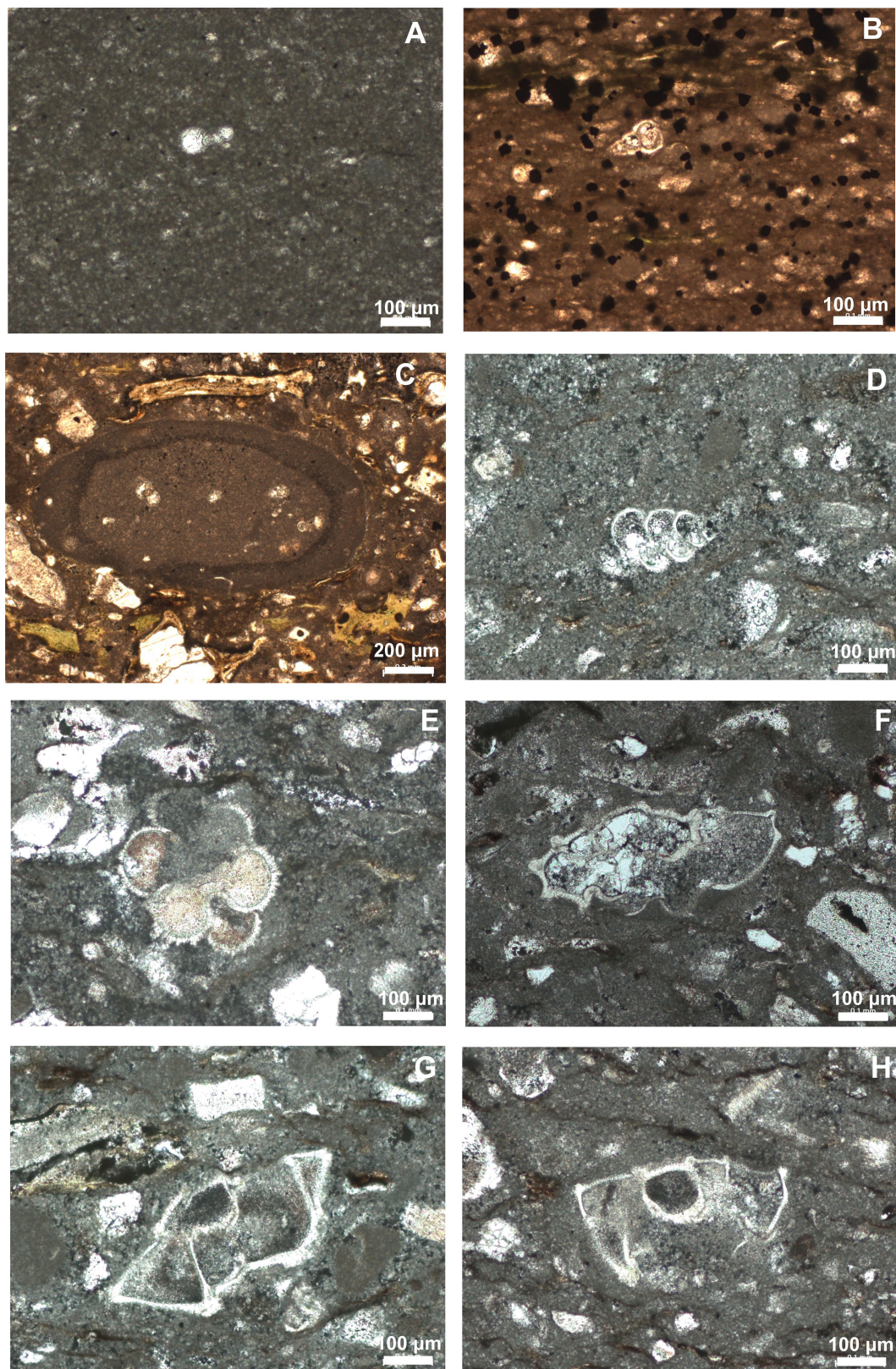




**Extended Data Fig. 2 | Trace fossils in core 40 section 1 of IODP hole M0077A.** Discrete burrows in the upper transitional unit and the lower limestone are circled and labelled by the genus. Above the base

of the limestone, trace fossils are abundant; representative examples are highlighted in the lower 10 cm of this interval. Ch, *Chondrites*; Pl, *Planolites*; Pa, *Palaeophycus*.

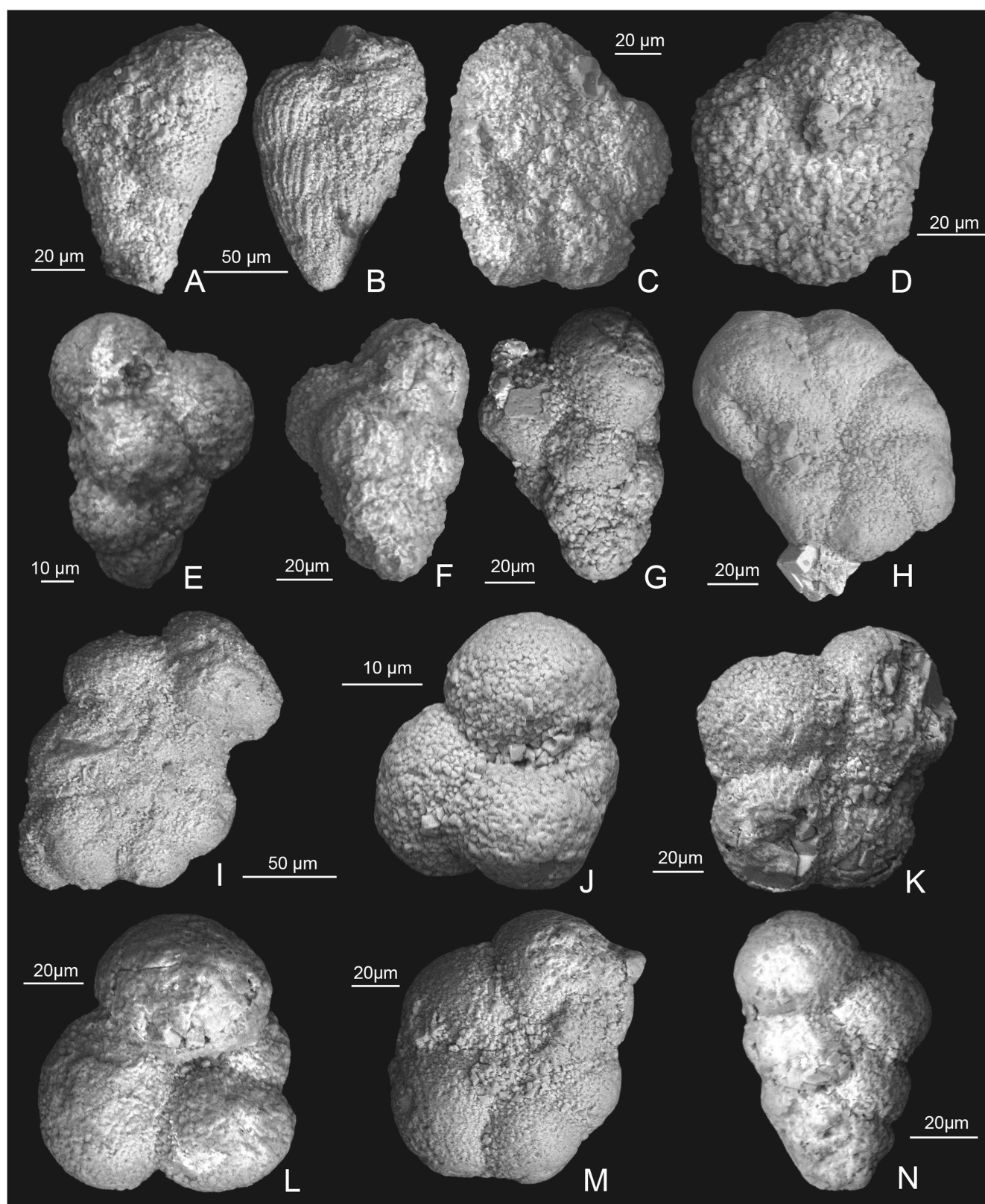




**Extended Data Fig. 3 | Reworked Cretaceous foraminifera in the transitional unit.** **a**, *Globigerinelloides* sp., sample 364-M0077A-40R-1-W, 55–56 cm. **b**, *Heterohelix* sp., sample 364-M0077A-40R-1-W, 104–105 cm. **c**, Clast of pelagic limestone containing older Cretaceous planktic foraminifera, sample 364-M0077A-40R-1-W, 106–110 cm. **d**, *Praegublerina pseudotessera*, sample 364-M0077A-40R-1-W, 118–129 cm.

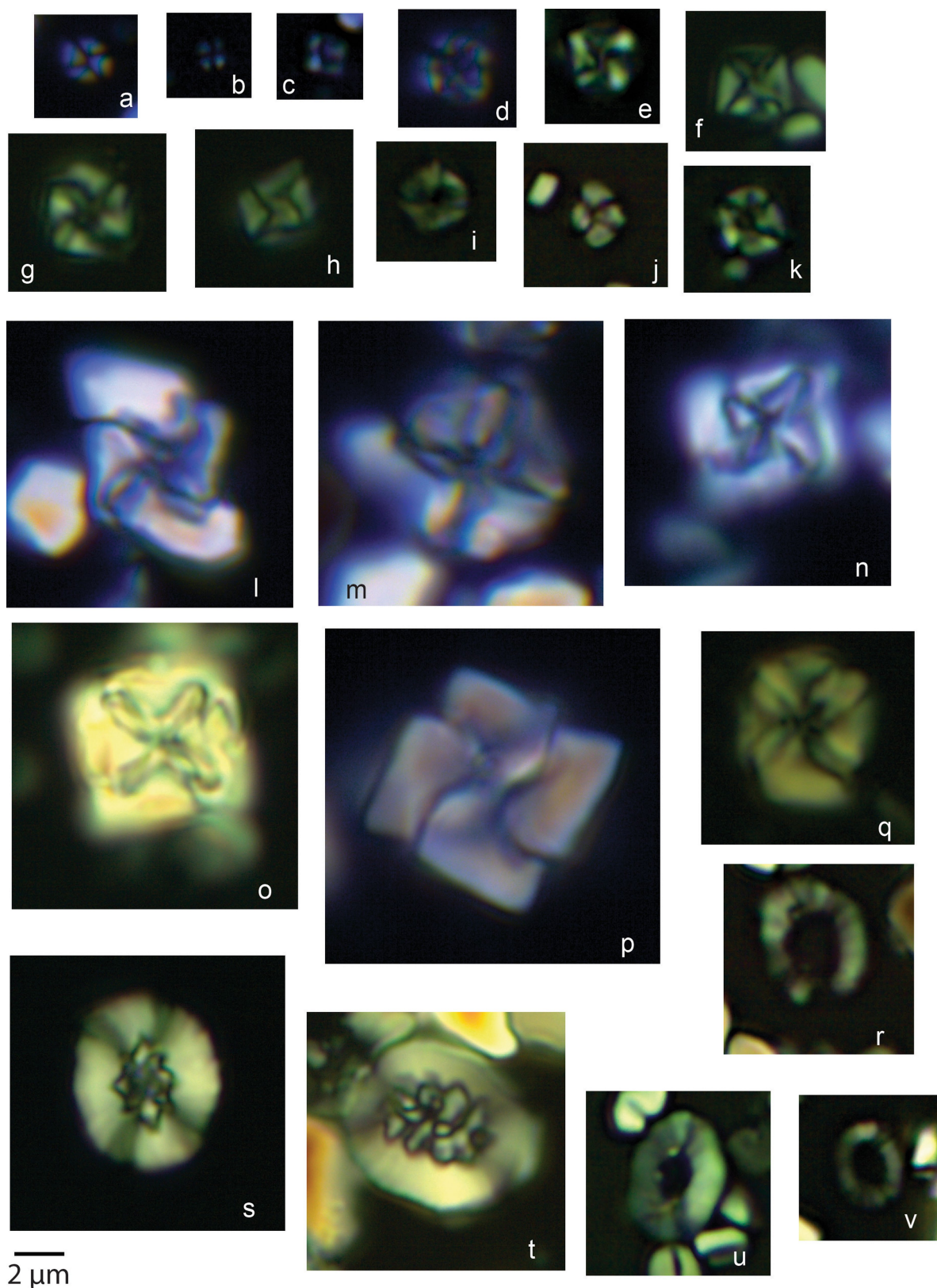
**e**, *Racemiguembelina powelli*, sample 364-M0077A-40R-1-W, 118–129 cm. **f**, *Globotruncana bulloides*, sample 364-M0077A-40R-1-W, 110–118 cm. **g**, *Globotruncanita stuartiformis*, sample 364-M0077A-40R-1-W, 118–129 cm. **h**, *Globotruncanita elevata*, sample 364-M0077A-40R-1-W, 118–129 cm. Scale bars, 100 μm.





**Extended Data Fig. 4 | Scanning electron micrographs of planktic foraminifera from core 40.** **a, b**, Examples of common reworked Cretaceous biserials, sample 364-M0077A-40R-1, 102–103 cm. **c**, *Muricohedbergella monmouthensis*, sample 364-M0077A-40R-1-W, 102–103 cm. **d**, *Muricohedbergella holmdelensis*, sample 364-M0077A-40R-1-W, 44–45 cm. **e**, *Guembelitra cretacea*, sample 364-M0077A-40R-1-W, 44–45 cm. **f**, *G. cretacea*, sample 364-M0077A-40R-1-W, 29–30 cm. **g**, *G. cretacea*, sample 364-M0077A-40R-1-W, 29–30 cm. **h**,

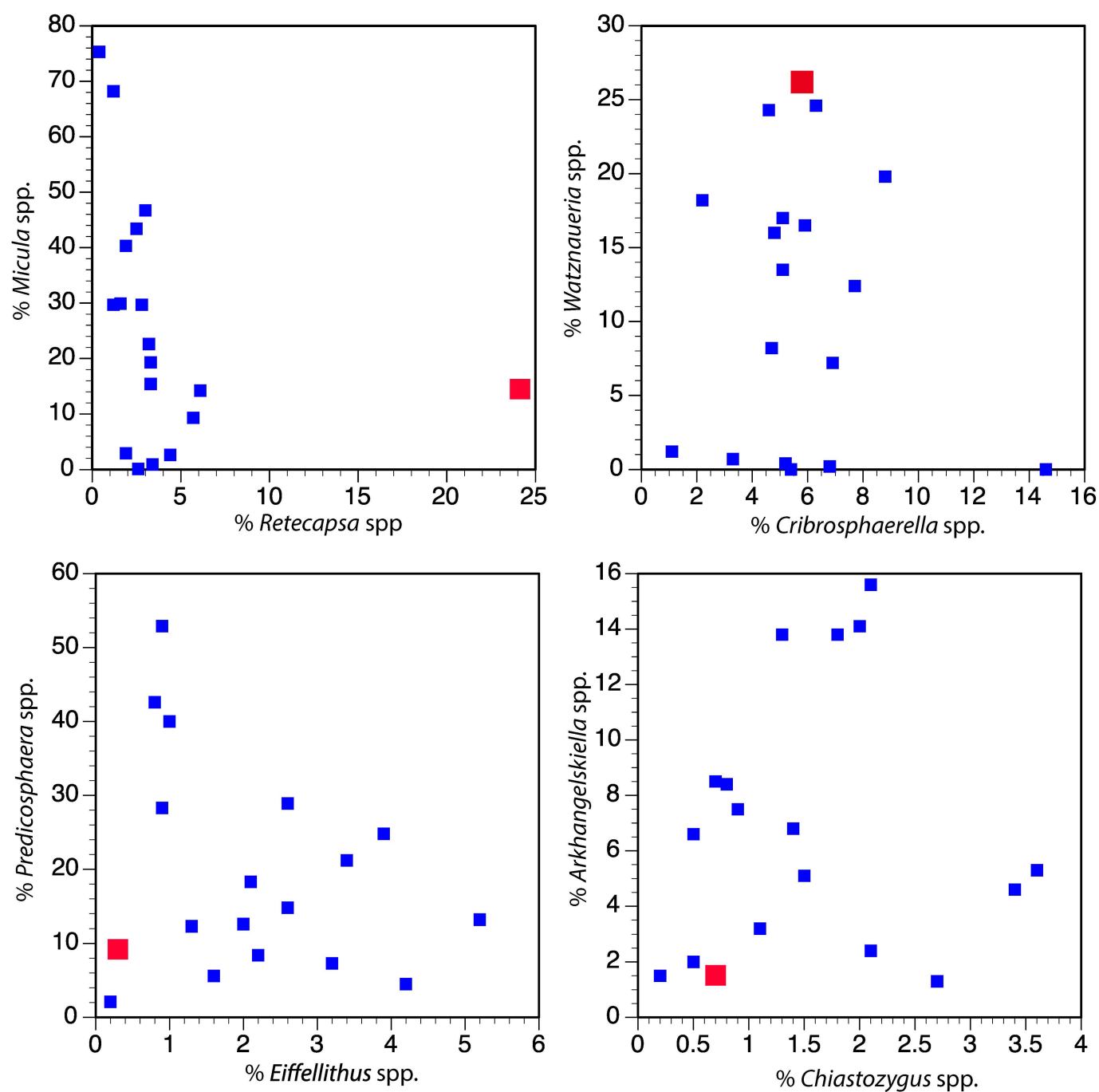
*Parvularugoglobigerina eugubina* 364-M0077A-40R-1-W, 31–32 cm. **i**, *P. eugubina*, sample 364-M0077A-40R-1-W, 31–32 cm. **j**, *Globoconusa daubjergensis*, sample 364-M0077A-40R-1-W, 31–32 cm. **k**, *Eoglobigerina eobulloides*, sample 364-M0077A-40R-1-W, 29–30 cm. **l**, *Eoglobigerina edita*, sample 364-M0077A-40R-1-W, 29–30 cm. **m**, *Praemurica taurica*, sample 364-M0077A-40R-1-W, 10–11 cm. **n**, *Chiloguembelina morsei*, sample 364-M0077A-40R-1-W, 10–11 cm.



**Extended Data Fig. 5 | Small and regular-sized nannofossils in the transitional unit.** All photographs from core 364-M0077-40R-1-W. Measurements in centimetres refer to depth in section 1 of core 40. **a–k**, Images of small *Micula* spp.: **a**, 55–56 cm; **b**, 41–42 cm; **c**, 95–96 cm; **d**, 41–42 cm; **e**, 90–91 cm; **f**, 94–95 cm; **g**, 91–92 cm; **h**, 91–92 cm; **i**, 45–46 cm;

**j**, 100–101 cm; **k**, 81–82 cm. **l–q**, Images of regular-sized *Micula* spp.: **l**, 44–45 cm; **m**, 41–42 cm; **n**, 51–52 cm; **o**, 105–106 cm; **p**, 97–98 cm; **q**, 36–37 cm. **s, t**, Images of regular-sized *Retecapsa* spp.: **s**, 85–86 cm; **t**, 100–101 cm. **r–v**, Images of small *Retecapsa* spp.: **r**, 100–101 cm; **u**, 71–72 cm, **v**, 100–101 cm. Scale bar, 2  $\mu$ m.





**Extended Data Fig. 6 | Relative abundances of major Maastrichtian calcareous nannoplankton.** Small blue squares are Maastrichtian sites from a global compilation<sup>12</sup>; larger red squares are from the transitional

unit at site M0077. These data demonstrate the unusual abundance of *Watznaueria* and *Retecapsa* at site M0077.

Extended Data Table 1 |  $^3\text{He}$  data

	start	stop	$^3\text{He}$	$^4\text{He}$	Absolute	Fraction	Maximum $^3\text{He}$ -Based
Sample	cm	cm	pcc/g	ncc/g	$^3\text{He}/^4\text{He}$	$^3\text{He}$ ET	Model Age (kyr)
KT39	39	40	0.0068	13.6	5.04E-07	0.96	6.0
KT48	48	49	0.0055	35.4	1.56E-07	0.87	4.9
KT59	59	60	0.0064	23.1	2.78E-07	0.92	4.0
KT68	68	69	0.0042	31.6	1.33E-07	0.84	2.9
KT79	79	80	0.0036	18.3	1.99E-07	0.9	1.9
KT89	89	90	0.0105	34.7	3.04E-07	0.93	0.9
KT99	99	100	0.0045	64.3	6.99E-08	0.70	0.1
KT106.5	107	108	0.0109	327	3.32E-08	0.37	0.0

# Parallel emergence of stable and dynamic memory engrams in the hippocampus

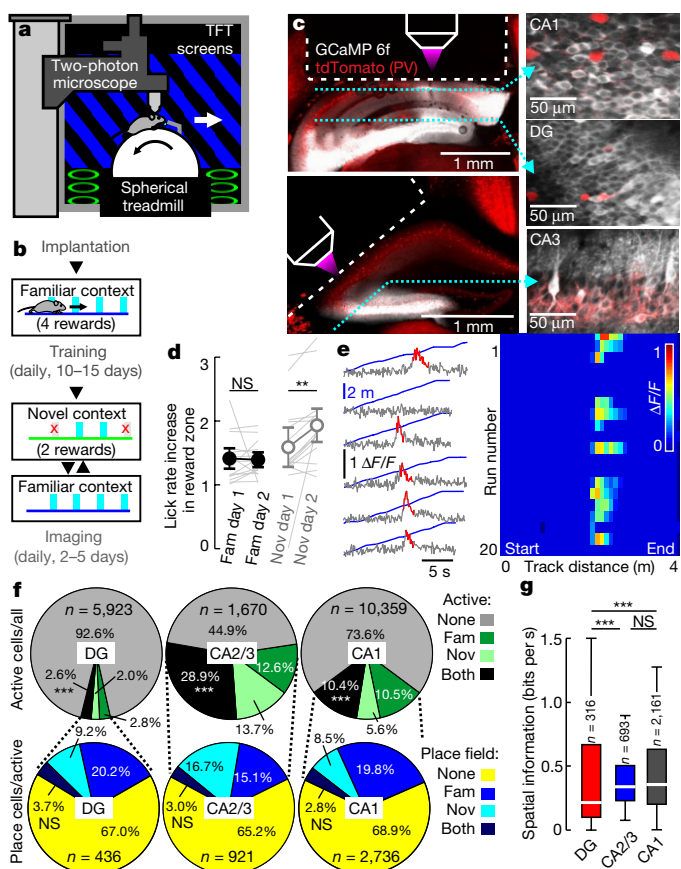
Thomas Hainmueller<sup>1,2,3</sup> & Marlene Bartos<sup>1\*</sup>

During our daily life, we depend on memories of past experiences to plan future behaviour. These memories are represented by the activity of specific neuronal groups or ‘engrams’<sup>1,2</sup>. Neuronal engrams are assembled during learning by synaptic modification, and engram reactivation represents the memorized experience<sup>1</sup>. Engrams of conscious memories are initially stored in the hippocampus for several days and then transferred to cortical areas<sup>2</sup>. In the dentate gyrus of the hippocampus, granule cells transform rich inputs from the entorhinal cortex into a sparse output, which is forwarded to the highly interconnected pyramidal cell network in hippocampal area CA3<sup>3</sup>. This process is thought to support pattern separation<sup>4</sup> (but see refs. 5,6). CA3 pyramidal neurons project to CA1, the hippocampal output region. Consistent with the idea of transient memory storage in the hippocampus, engrams in CA1 and CA2 do not stabilize over time<sup>7–10</sup>. Nevertheless, reactivation of engrams in the dentate gyrus can induce recall of artificial memories even after weeks<sup>2</sup>. Reconciliation of this apparent paradox will require recordings from dentate gyrus granule cells throughout learning, which has so far not been performed for more than a single day<sup>6,11,12</sup>. Here, we use chronic two-photon calcium imaging in head-fixed mice performing a multiple-day spatial memory task in a virtual environment to record neuronal activity in all major hippocampal subfields. Whereas pyramidal neurons in CA1–CA3 show precise and highly context-specific, but continuously changing, representations of the learned spatial sceneries in our behavioural paradigm, granule cells in the dentate gyrus have a spatial code that is stable over many days, with low place- or context-specificity. Our results suggest that synaptic weights along the hippocampal trisynaptic loop are constantly reassigned to support the formation of dynamic representations in downstream hippocampal areas based on a stable code provided by the dentate gyrus.

To study hippocampal memory engrams during long-term learning, we designed a goal-oriented learning task for head-fixed mice. Mice ran on a spherical treadmill to collect soy milk rewards on a 4-m-long virtual linear track displayed on monitors around the animal. After at least 10 days of familiarization to this track (familiar context), imaging sessions started in which mice ran alternately on this familiar and a visually different, novel track with different reward sites (Fig. 1a, b, Supplementary Video 1). Animals consistently licked more often inside than outside reward zones on both tracks (Fig. 1d). Initially, overall licking and reward-related licking were lower in the novel context than in the familiar context (Extended Data Fig. 1c, d). These differences vanished with learning. On the novel track, the ratio of rewarded to erroneous licks increased markedly on the second training day (Fig. 1d), indicating that mice remembered the rewarded locations.

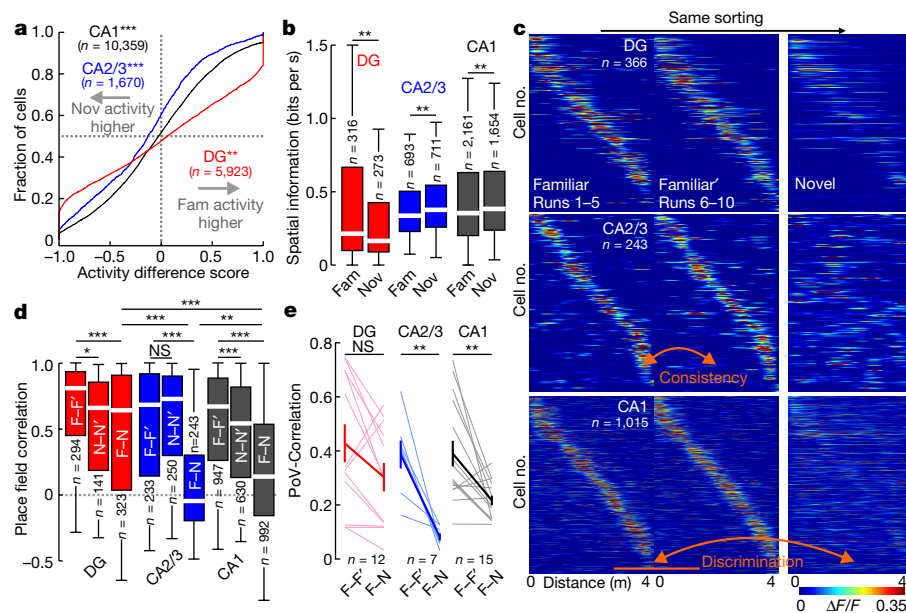
To measure hippocampal neuronal activity, mice were injected with adeno-associated viruses designed to express the fluorescent calcium indicator GCaMP6f pan-neuronally in CA1 and the dentate gyrus (DG) or CA3 (Fig. 1c). A chronic transcortical imaging window was implanted above CA1 to perform two-photon imaging of CA1 or DG neurons<sup>13</sup> (Supplementary Video 2). Implantation did not

impair spatial learning in a Barnes maze (Extended Data Fig. 1h, i). CA1 and DG neurons were imaged at depths of around 150  $\mu\text{m}$  and around 700  $\mu\text{m}$ , respectively. To image CA3, we implanted a more



**Fig. 1 | Two-photon calcium imaging of hippocampal place cell activity in a virtual environment.** **a**, Experimental schematic. **b**, Behaviour timeline (see Methods). **c**, Left, CA1 and DG (top) and CA2/3 (bottom) implantation sites. GCaMP6f (white) and tdTomato (tdT; red) in parvalbumin (PV)-expressing interneurons. Dotted lines, imaging planes. Right, GCaMP6f and tdT fluorescence in vivo. **d**, Ratio between rewarded and non-rewarded licks in the familiar (fam, filled circles) and novel (nov, open circles) contexts ( $n = 15$  mice; two-sided signed rank-sum test). **e**, Calcium traces (grey) with significant transients (red; see Methods) of a GC and linear-track position (blue) over time. Dotted lines, imaging planes. Right, calcium activity over track distance of the same GC. **f**, Top, fraction of active (more than 0.05 transients per s) cells among all neurons. Test for population overlap ( $\chi^2$  test). Bottom, cells with place fields among active cells. **g**, Spatial information for all familiar-track-active neurons (ANOVA on ranks, Dunn's test). Boxes, 25th to 75th percentiles; bars, median; whiskers, 99% range. NS, not significant; \*\*\* $P < 0.001$ . Error bars denote s.e.m. For exact  $P$  values see Supplementary Table 1.

<sup>1</sup>Institute for Physiology I, Systemic and Cellular Neurophysiology, University of Freiburg, Freiburg, Germany. <sup>2</sup>Spemann Graduate School of Biology and Medicine (SGBM), University of Freiburg, Freiburg, Germany. <sup>3</sup>Faculty of Biology, University of Freiburg, Freiburg, Germany. \*e-mail: marlene.bartos@physiologie.uni-freiburg.de



**Fig. 2 | Pyramidal cells in CA1 and CA2/3 discriminate between contexts.** **a**, Activity difference scores (see Methods) between novel and familiar contexts for all cells (two-sided signed rank-sum test: novel versus familiar activity). **b**, Mean spatial information of active cells (two-sided rank-sum test). **c**, Familiar-track place cell activity plotted for the first (left) and second (middle) block of familiar-context runs and for the novel context (right). **d**, Mean activity correlations of place cells within familiar

(left) and novel context (middle) runs and between contexts (right bars; ANOVA on ranks, Dunn's test). **e**, PoV correlations (see Methods) for all experiments within familiar-context runs and between contexts (thin lines). Thick lines denote mean  $\pm$  s.e.m. (two-sided paired *t*-test). Boxes, 25th to 75th percentiles; bars, median; whiskers, 99% range. \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001; NS, not significant. For exact *P* values see Supplementary Table 1.

lateral window<sup>14</sup> (Fig. 1c). Data were obtained predominantly from CA3 (Extended Data Fig. 2d), but some CA2 cells may also have been included<sup>14</sup>. In all cases, we used fast volumetric scanning to simultaneously record about 500 neurons (see Methods, Supplementary Videos 3–5).

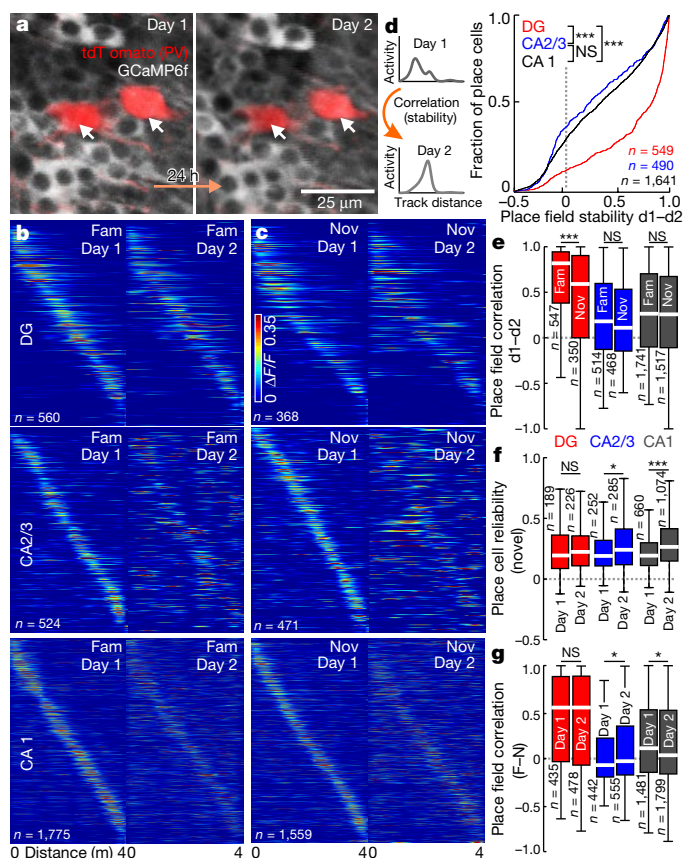
We first analysed neuronal activity in the familiar and novel contexts (Fig. 1). Consistent with previous findings<sup>11,12,15,16</sup>, pyramidal cells (PYRs) in CA1–CA3 were substantially more active than granule cells (GCs) (Fig. 1f, Extended Data Fig. 3a, b). We also determined the fraction of cells that was active in the familiar, novel or both contexts with more than 0.05 calcium transients per second. Activation of hippocampal neurons might be predetermined by intrinsic properties<sup>17</sup>. In line with this idea, we found a marked overlap of active neuronal ensembles between contexts (Fig. 1f, upper row). About 35% of these active neurons had a clearly defined place field (see Methods) on the first recording day in either the novel or the familiar context, or both (Fig. 1f, lower row). Because many neurons were active in both contexts, we investigated whether active neurons were more likely to have a place field in both contexts. However, the familiar- and novel-context place cells appeared to form independent subgroups within the active cell population (Fig. 1f, lower row) indicating that a separate place-coding group or 'engram' might exist for each context. Further comparison of spatial coding properties revealed lower average spatial information (see Methods) per active cell (Fig. 1g) and wider place fields (Extended Data Fig. 3c) in GCs compared to CA1–3 PYRs. Thus, GC activity is sparse and has broader and less precise spatial tuning than PYR activity.

Next, we compared neuronal activity between contexts (Fig. 2). Consistent with a previous study<sup>15</sup> and their inputs from the entorhinal cortex<sup>18</sup>, mean activity in GCs decreased in the novel context, whereas mean activity of CA1 and CA2/3 PYRs increased (Fig. 2a). Similarly, novel-context spatial information was markedly lower in the DG, but higher in both CA regions (Fig. 2b). Additionally, there was a trend towards higher place cell numbers on the familiar track than on the novel track, particularly in the DG (*n* = 30.50 versus 16.42 place cells per experiment, 12 experiments, *P* = 0.066, paired *t*-test; Extended Data Fig. 3f). We next investigated the cause of these activity differences between contexts. Hippocampal  $\gamma$ -aminobutyric acid (GABA)

interneurons contribute to separation of memory engrams<sup>19</sup> and formation of place fields<sup>13</sup>. We therefore analysed the activity of parvalbumin (PV)-expressing interneurons (PVIs; Extended Data Fig. 4), the most abundant subtype of interneurons in the hippocampus. PVI activity in CA1 and the DG correlated positively with running speed<sup>13,20</sup> (Extended Data Fig. 4c–h). PVIs in the DG, but not in CA1, showed decreased activity in the novel context (Extended Data Fig. 4i–l), contrasting with reports from unidentified DG interneurons<sup>15</sup>. Thus, our data argue against suppression of GCs by enhanced PVI activity, and are instead consistent with predominant recruitment of DG PVIs by local GC inputs.

To probe neuronal discrimination between contexts, we first determined the consistency of place cell firing on the same track between the first and the second block of five consecutive runs. Place cell consistency in the familiar context was high in all hippocampal subfields (Fig. 2c, d; F–F'). The same measure and trial-to-trial reliability were generally lower for novel-context runs, indicating an initially less reliable representation (Fig. 2c, d; N–N'; Extended Data Fig. 3i). Next, we quantified place cell remapping between the familiar and novel contexts. Unexpectedly, activity map correlations between contexts were substantially higher for DG place cells than in CA1 and CA2/3 (Fig. 2c, d; F–N). We confirmed this finding separately in two mice by imaging neuronal activity in CA1 and DG of the same mice (Extended Data Fig. 5). Thus, DG place cell activity was similar between contexts, whereas that of CA1–3 place cells was highly discriminative. We also calculated population vectors (PoVs) for both contexts from the mean calcium activity maps of all cells. PoVs were significantly more dissimilar between contexts as compared to independent runs within the familiar context in CA1 and CA2/3 (*P* = 0.004, both regions, paired *t*-test), but not in the DG (*P* = 0.051, Fig. 2e). Indeed, activity map correlations between contexts were markedly lower in CA1 and CA2/3 than in the DG, indicating stronger remapping in CA1–3. GCs might encode travelled distance and therefore show low context-selectivity. To test this possibility, we let mice run on a simplified linear track with striped walls but no further contextual information (Extended Data Fig. 6a). Under these conditions, GC activity and spatial information were low, and GCs did not show consistent place fields (Extended Data Fig. 6b, c),





**Fig. 3 | GC place fields are highly stable across days.** **a**, Illustrative example of CA1 cells imaged on subsequent days. **b**, Activity of familiar-track place cells sorted for day 1. **c**, As in **b** for novel-track place cells. **d**, Left, experimental schematic. Right, activity map correlations between days for all day 1 (d1) place cells (ANOVA on ranks, Dunn's test). **e**, Activity map correlations between days for familiar-context (left) and novel-context (right) place cells. **f**, Mean trial-to-trial reliability of novel-track place cell responses on days 1 and 2. **g**, Activity map correlations between contexts of day 1 (left) and day 2 (right) place cells. **e–g**, Two-sided rank-sum test. Boxes, 25th to 75th percentiles; bars, median; whiskers, 99% range. \* $P < 0.05$ , \*\*\* $P < 0.001$ ; NS, not significant. For exact  $P$  values see Supplementary Table 1.

indicating that they encode the general task layout rather than mere distance. Thus, GCs show reliable place representations and low context discrimination, whereas CA2/3 PYRs and CA1 PYRs prominently encode contextual differences.

To investigate place field stability throughout learning, we imaged the same cells in both contexts on two subsequent days (Fig. 3, Extended Data Fig. 7). Whereas GCs maintained their place field locations in the same context, CA1 PYRs and CA2/3 PYRs displayed substantial remapping (Fig. 3b–d). This was characterized by lower activity map correlations (Fig. 3d, e) and larger shifts of the preferred firing location (Extended Data Fig. 7c). Despite the generally high GC place field stability, activity map correlations between days were lower for GCs encoding the novel context than for those encoding the familiar context. By contrast, hippocampal PYRs showed similarly low stability in both contexts (Fig. 3e). Thus, GCs have stable place fields, whereas place fields in other hippocampal subfields change over days.

Place cell stability in CA1 may depend on environmental complexity<sup>9</sup>. We therefore repeated our experiments in a virtual context without distal visual cues ('poor') and a highly enriched, multi-sensory track ('rich'; Supplementary Video 6). Notably, the number of place cells was similar between all tracks, but their firing rate, spatial information and day-to-day stability were markedly reduced on the 'poor' track (Extended Data Fig. 8). We observed no differences between the 'rich' track and our standard contexts, indicating that

CA1 place cell representations are also dynamic over days in complex environments.

Next, we investigated learning-induced changes in spatial coding. From day 1 to day 2, there was a substantial increase in the trial-to-trial reliability of place cells in CA1–3, but not in the DG (Fig. 3f, Extended Data Fig. 7d). The DG is required for context discrimination<sup>5,6</sup>. We therefore tested whether neuronal activity became more distinct between contexts with learning. Unexpectedly, activity correlations between contexts were unchanged in GCs from day 1 to day 2 but were lower for CA1 place cells on day 2 and remained negative in CA2/3 (Fig. 3g). Thus, improved behavioural context discrimination was accompanied by a decorrelation of place cell activity in CA1, but not in the DG.

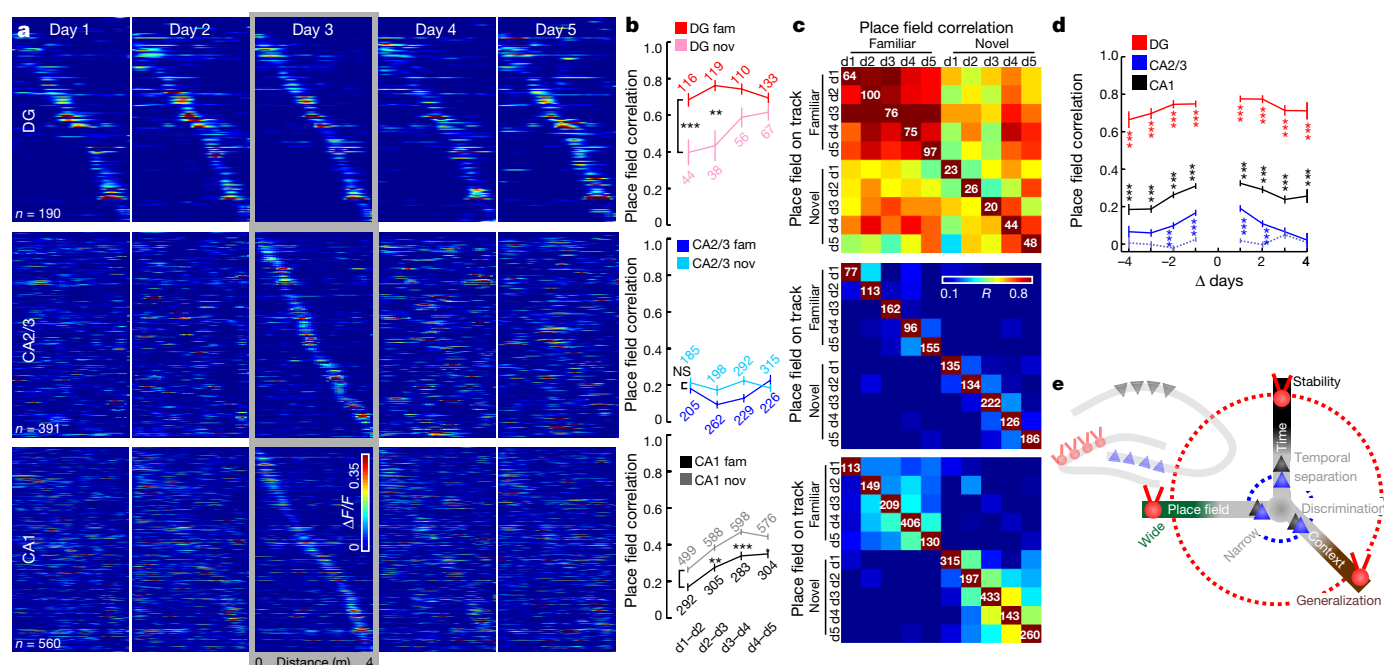
In light of recent findings<sup>20</sup>, we compared place coding between the deep and superficial sublayers of CA1. Notably, place field stability across days was slightly higher in deep-layer PYRs (45% difference,  $P = 0.046$ ; Extended Data Fig. 9i), albeit at generally low levels. Place field reliability and context discrimination were comparable between sublayers (Extended Data Fig. 9g, h).

To investigate the development of spatial representations over the time-course of hippocampus-dependent memory<sup>2</sup>, we continued imaging sessions for five subsequent days (Fig. 4, Extended Data Fig. 10). Whereas the number of place cells was similar for each context and day (Fig. 4c, white numbers), their firing locations changed markedly in some hippocampal sub-areas. Familiar-context place fields of GCs remained stable throughout all days (Fig. 4a–c) and novel-context place cells showed gradually increasing stability (Fig. 4b, c, Extended Data Fig. 10b). By contrast, CA1 and CA2/3 activity in the same contexts became rapidly more dissimilar over days. For CA2/3 cells, activity map correlations over more than two days dropped below chance levels (Fig. 4d), demonstrating that these neurons constantly remap their place fields.

Dynamic coding has been described in CA1<sup>7–9,21</sup>, CA2<sup>10</sup> and other associative areas<sup>22,23</sup>. A gradual variation of active CA1 ensembles links contextual memories acquired close in time<sup>7,8,21</sup> and remapping of individual PYRs is driven by synaptic plasticity<sup>13,24</sup>. By contrast, neuronal ensemble activity in motor areas stabilizes throughout learning<sup>25</sup>. In the hippocampus, temporally stable coding of GCs may induce heterosynaptic plasticity at CA3 PYR dendrites by associating their activity with temporally dynamic inputs from the entorhinal cortex<sup>23</sup> or other CA3 PYRs<sup>26</sup>. This hypothesis would explain why CA3 ensembles can trigger memory recall independent of GC input even when the DG is required for initial task learning<sup>27,28</sup>. Our results, together with previous findings<sup>8,10</sup>, indicate that CA3 coding can be dynamic or stable, potentially depending on the behaviour, proximo-distal location within CA3 (Extended Data Fig. 2e), virtual versus real-world navigation or species differences in entorhinal cortex innervation<sup>29</sup>. When CA3 is stable, a mechanism similar to that described above may apply at CA3–CA1 synapses.

Traditionally, similar memories are thought to be represented by largely non-overlapping populations of GCs<sup>4</sup>. However, recent findings indicate that GCs remap only between widely dissimilar environments<sup>11</sup>, while other cell types (for example, mossy cells) discriminate between similar contexts<sup>12</sup>. Accordingly, CA2/3 PYR activity was most discriminative between our virtual contexts (Fig. 2d, e). The high similarity of our—mostly mature<sup>6</sup>—GC activity between contexts may explain why mature GCs mediate generalization between similar contexts rather than pattern separation<sup>5</sup>.

Our results further suggest that the hippocampus combines stable and dynamic coding and reunites findings of temporally varying neuronal ensembles encoding the same environment<sup>7,8</sup> with reports of stable behavioural output upon DG engram reactivation over weeks<sup>2,27</sup>. Given that the DG is required for the extinction or modification of existing memories acquired in the same scenery<sup>28,30</sup>, our data support the hypothesis that GCs provide a simplistic but stable representation of the global environment<sup>11,12</sup> that serves as a blueprint for spatially and contextually precise, but temporally varying, CA1–3 engrams



**Fig. 4 | Stable coding of GCs persists throughout multiple days, whereas CA2/3 PYRs constantly remap over time.** **a**, Activity maps of all familiar-context place cells sorted by day 3. **b**, Development of activity map correlations between subsequent days for familiar- (dark) and novel-context (light) place cells (numbers show  $n$ ; ANOVA on ranks, Dunn's test; mean  $\pm$  s.e.m.). **c**, Mean activity map correlations (colour coded; Pearson's  $R$ ) over 5 days and two contexts as indicated on the x-axis. Each row shows mean correlation values for cells that had a place field on the day and track indicated on the y-axis (white numbers show  $n$ ). **d**, Mean activity map

correlations for familiar-context place cells over days passed. Blue dotted line, chance level correlations for CA2/3 cells obtained by shuffling cell IDs (two-sided rank-sum test, actual versus shuffled correlations, Bonferroni correction; mean  $\pm$  s.e.m.). **e**, Schematic: GCs show a highly stable environment representation with low spatial and context selectivity. By contrast, PYRs form highly context-, place- and time-specific ensembles. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ . For exact  $P$  and  $n$  values in **d** see Supplementary Table 1.

(Fig. 4e). Such an encoding scheme would enable one to associate memories acquired in the same global environment but still to discriminate between slightly different or temporally separate instances of these memories.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0191-2>.

Received: 20 November 2017; Accepted: 30 April 2018;

Published online 6 June 2018.

- Ramirez, S. et al. Creating a false memory in the hippocampus. *Science* **341**, 387–391 (2013).
- Kitamura, T. et al. Engrams and circuits crucial for systems consolidation of a memory. *Science* **356**, 73–78 (2017).
- Pernia-Andrade, A. J. & Jonas, P. Theta-gamma-modulated synaptic currents in hippocampal granule cells in vivo define a mechanism for network oscillations. *Neuron* **81**, 140–152 (2014).
- Chawla, M. K. et al. Sparse, environmentally selective expression of Arc RNA in the upper blade of the rodent fascia dentata by brief spatial experience. *Hippocampus* **15**, 579–586 (2005).
- Nakashiba, T. et al. Young dentate granule cells mediate pattern separation, whereas old granule cells facilitate pattern completion. *Cell* **149**, 188–201 (2012).
- Danielson, N. B. et al. Distinct contribution of adult-born hippocampal granule cells to context encoding. *Neuron* **90**, 101–112 (2016).
- Rubin, A., Geva, N., Sheintuch, L. & Ziv, Y. Hippocampal ensemble dynamics timestamp events in long-term memory. *eLife* **4**, e12247 (2015).
- Mankin, E. A. et al. Neuronal code for extended time in the hippocampus. *Proc. Natl Acad. Sci. USA* **109**, 19462–19467 (2012).
- Kentros, C. G., Agnihotri, N. T., Streater, S., Hawkins, R. D. & Kandel, E. R. Increased attention to spatial context increases both place field stability and spatial memory. *Neuron* **42**, 283–295 (2004).
- Mankin, E. A., Diehl, G. W., Sparks, F. T., Leutgeb, S. & Leutgeb, J. K. Hippocampal CA2 activity patterns change over time to a larger extent than between spatial contexts. *Neuron* **85**, 190–201 (2015).
- GoodSmith, D. et al. Spatial representations of granule cells and mossy cells of the dentate gyrus. *Neuron* **93**, 677–690.e5 (2017).
- Senzai, Y. & Buzsáki, G. Physiological properties and behavioral correlates of hippocampal granule cells and mossy cells. *Neuron* **93**, 691–704.e5 (2017).
- Sheffield, M. E. J., Adoff, M. D. & Dombeck, D. A. Increased prevalence of calcium transients across the dendritic arbor during place field formation. *Neuron* **96**, 490–504.e5 (2017).
- Rajasethupathy, P. et al. Projections from neocortex mediate top-down control of memory retrieval. *Nature* **526**, 653–659 (2015).
- Nitz, D. & McNaughton, B. Differential modulation of CA1 and dentate gyrus interneurons during exploration of novel environments. *J. Neurophysiol.* **91**, 863–872 (2004).
- Leutgeb, J. K., Leutgeb, S., Moser, M.-B. & Moser, E. I. Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* **315**, 961–966 (2007).
- Diamantaki, M., Frey, M., Berens, P., Preston-Ferrer, P. & Burgalossi, A. Sparse activity of identified dentate granule cells during spatial exploration. *eLife* **5**, e20252 (2016).
- Burgalossi, A., von Heimendahl, M. & Brecht, M. Deep layer neurons in the rat medial entorhinal cortex fire sparsely irrespective of spatial novelty. *Front. Neural Circuits* **8**, 74 (2014).
- Stefanelli, T., Bertolini, C., Lüscher, C., Müller, D. & Mendez, P. Hippocampal somatostatin interneurons control the size of neuronal memory ensembles. *Neuron* **89**, 1074–1085 (2016).
- Lee, S.-H. et al. Parvalbumin-positive basket cells differentiate among hippocampal pyramidal cells. *Neuron* **82**, 1129–1144 (2014).
- Cai, D. J. et al. A shared neural ensemble links distinct contextual memories encoded close in time. *Nature* **534**, 115–118 (2016).
- Driscoll, L. N., Pettit, N. L., Minderer, M., Chetih, S. N. & Harvey, C. D. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell* **170**, 986–999.e16 (2017).
- Tsao, A. et al. Integrating time in the entorhinal cortex. *Society for Neuroscience* 084.21.2017 (2017).
- Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).
- Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).
- Brandalise, F. & Gerber, U. Mossy fiber-evoked subthreshold responses induce timing-dependent plasticity at hippocampal CA3 recurrent synapses. *Proc. Natl Acad. Sci. USA* **111**, 4303–4308 (2014).
- Roy, D. S. et al. Memory retrieval by activating engram cells in mouse models of early Alzheimer's disease. *Nature* **531**, 508–512 (2016).
- Bernier, B. E. et al. Dentate gyrus contributes to retrieval as well as encoding: evidence from context fear conditioning, recall, and extinction. *J. Neurosci.* **37**, 6359–6371 (2017).

29. van Groen, T., Miettinen, P. & Kadish, I. The entorhinal cortex of the mouse: organization of the projection to the hippocampal formation. *Hippocampus* **13**, 133–149 (2003).
30. Kheirbek, M. A. et al. Differential control of learning and anxiety along the dorsoventral axis of the dentate gyrus. *Neuron* **77**, 955–968 (2013).

**Acknowledgements** We thank H.-J. Weber, C. Paun and C. Schmidt-Hieber for advice and help with setting up the virtual environment system; K. Winterhalter and K. Semmler for technical support; and J. Sauer, M. Strueber and M. Eyre for comments on earlier versions of the manuscript. This work was funded by the German Research Foundation (FOR2143, M.B.) and ERC-AdG 787450 (M.B.). This work was supported in part by the Excellence Initiative of the German Research Foundation (GSC-4, Spemann Graduate School; T.H.).

**Reviewer information** *Nature* thanks M. Brecht and S. Leutgeb for their contribution to the peer review of this work.

**Author contributions** T.H. and M.B. conceived the study, designed the experiments and wrote the manuscript. T.H. performed experiments and analysed data.

**Competing interests** The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0191-2>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0191-2>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to M.B.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## METHODS

**Mice.** All experiments involving animals were carried out according to national and institutional guidelines and approved by the 'Tierversuchskommission' of the Regierungspräsidium Freiburg (license no. G16/037) in accordance with national legislation. *B6;129P2-Pvalb<sup>tm1.1(cre)Arb</sup>/J* mice (PV-Cre; The Jackson laboratory) crossed with *B6.Cg-Gt(Rosa)26Sor<sup>tm9(CAG-tdTomato)Hze</sup>/J* mice (Ai9-reporter; The Jackson laboratory) were used for all experiments at an age of 7–15 weeks ( $n = 6$  for DG, 5 for CA2/3, 11 for CA1 and 7 for the 'poor–rich–novel' (Extended Data Figs. 6, 8) recordings). Age-matched C57/Bl6 mice injected with GFP virus were used as control group for Barnes maze experiments. Mice were housed on a 12-h light–dark cycle in groups of 2–5 mice. After the start of the post-window-implantation training and food restriction, mice were housed individually. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Virus injections and head plate implantation.** All surgical procedures were performed in a stereotactic apparatus (Kopf instruments) under anaesthesia with 1–2% isoflurane and analgesia using 0.1 mg kg<sup>-1</sup> buprenorphine. A small (~0.5–1 mm diameter) craniotomy was made over the hippocampus and 500 nl AAV1.Syn.GCaMP6f.WPRE.SV4 (titre  $4.65 \times 10^{13}$  vg (viral genomes) per ml; University of Pennsylvania Vector Core) were injected into CA3 (A/P –1.7 mm; M/L 1.9 mm; D/V –1.9 mm) or DG and CA1 (A/P –2.0 mm; M/L 2.0 mm; D/V –2.0 and/or –1.4 mm). GFP controls were injected with equal amounts of AAV1.CAG.GFP (titre  $3.2 \times 10^{12}$  vg per ml; University of North Carolina Vector core) with the same coordinates. In the same surgery session, mice were implanted with a stainless-steel head plate (25 × 10 × 0.8 mm with an 8-mm central aperture). The head plate was oriented horizontally for CA1 and DG imaging implantations and with a 20° lateral angle for CA3 imaging implantations. Mice were allowed to recover from surgery for at least 5 days before training sessions commenced. Postoperative analgesic treatment was continued with carprofen (5 mg kg<sup>-1</sup> body weight) for 3 days after surgery.

**Imaging window implantation.** Cortical excavation and imaging window implantation were performed more than 10 days after the initial virus injection, according to published protocols<sup>14,31</sup>. A craniotomy (diameter 3 mm) was made centred at A/P –1.5 mm, M/L –1.5 mm for CA1/DG imaging and A/P –1.5 mm, M/L –2.5 mm for CA3 recordings. For implantations over CA3, the head of the mouse was tilted by 20°, so that the implantation plane was parallel to the lateral part of the pyramidal cell layer. Parts of the somatosensory cortex and posterior parietal association cortex were gently aspirated while being irrigated with chilled saline. We continued aspiration until the external capsule was exposed. The outer part of the external capsule was then gently peeled away using fine forceps, leaving the inner capsule and the hippocampus itself undamaged. The imaging window implant consisted of a 3-mm diameter coverslip (CS-3R, Warner Instruments) glued to the bottom of a stainless steel cannula (3-mm diameter, 1.2–1.5-mm height). The window was gradually lowered into the craniotomy using forceps until the glass was in contact with the external capsule. The implant was then fixed to the skull using cyanoacrylate. Mice were allowed to recover from window implantation for 2–3 days.

**Barnes maze.** To test for potential detrimental effects of our implantation on normal hippocampal function, we compared animals implanted for our imaging experiments with GFP-injected control mice ( $n = 8$ , each) in a Barnes maze paradigm. The mice were tested on three consecutive days on a 1-m diameter Barnes maze (Noldus) with 20 equally spaced holes at the perimeter in a sound-isolated chamber with distal visual cues on the walls. One session with four runs interleaved by approximately 30 s was carried out per day. The location of the escape hole in relation to the visual cues was kept constant throughout experiments. The mouse was placed onto a starting point in a different quadrant of the maze for each trial. A trial ended after 2 min or when the mouse's body had completely entered the escape hole. Mice were tracked using video tracking software (Ethovision, Noldus) and the distance travelled (m) was analysed.

**Virtual environment setup and behavioural training.** Our custom virtual environment setup consisted of an air-supported polystyrene ball (20-cm diameter), similar to other published designs<sup>13,31–35</sup>. A small metal axle was attached to the side of the ball to constrain the ball motion to the forward–backward direction. The motion of the ball was monitored with an optical sensor (G-500, Logitech) and translated into forward motion through the virtual environment. The forward gain was adjusted so that 4 m of distance travelled along the circumference of the ball equalled one full traversal along the linear track. When the mouse had reached the end of the track, screens were blanked for 5–10 s and the mouse was 'teleported' back to the start of the linear track. The virtual environment was displayed on four TFT monitors (19" screen diagonal, Dell) arranged in a hexagonal arc around the mouse and placed ~25 cm away from the head, thereby covering ~260° of the horizontal and ~60° of the vertical visual field of the mouse, similar to ref.<sup>35</sup>. The virtual environment was created and simulated using the open-source 3D rendering

software Blender<sup>33</sup>. The track consisted of textured walls, floors and other 3D rendered objects at the tracks sides as visual cues (Extended Data Fig. 1g). Potential reward locations were marked with visual and acoustic cues, and 4 µl of soy milk was gradually dispensed through a spout in front of the mouse as long as the mouse waited in a rewarded location (see Supplementary Video 1). The simplified ('poor') track used for the experiments shown in Extended Data Fig. 8 was derived from the standard familiar context, but all rendered cues and objects were removed, except for a homogenous texture on the walls to both sides of the mouse. The context for the experiments in Extended Data Fig. 6 was essentially the same, but the first three reward locations were additionally removed to limit positional information to idiothetic cues alone. For the multisensory 'rich' context (Extended Data Fig. 8, Supplementary Video 6) a large number of 3D animated objects was added to the standard track. Furthermore, a brief (400-ms) 4-kHz tone was played as auditory background at 1 Hz repetition rate and a vanilla-scented piece of cardboard was placed into the behavioural setup. Finally, small cloth objects were attached to a foam disk (10-cm diameter), which was mounted on the axis of a stepper motor. The motor was controlled by an Arduino-board and programmed to move the objects into the range of the mouse's whiskers according to its position on the virtual track (Supplementary Video 6).

Five days after head plate implantation, mice were placed in the virtual environment for 10–30 min daily, with gradually increasing timespans. During this time, only the familiar context was available to the mice. After 4–5 days of habituation, mice showed consistent running and reward-related licking. The mice were thereafter implanted with the cortical window and allowed to recover from the surgery. After recovery, food scheduling was initiated with a goal of 85–90% of the ad libitum body weight. Simultaneously, training in the virtual environment was re-initiated for 30–60 min daily in the familiar context until consistent reward licking was observed in all animals and familiarization to the context had been achieved for at least 10 days in total before start of the imaging sessions.

From the first day of the imaging session, mice were introduced to a novel context, which had different visual cues and floor and wall textures, but had the same dimensions as the familiar context including the four marked reward locations. On the novel track, two of these reward sites were disabled (that is, the auditory cue was still given, but no reward was dispensed). Licking by the mice was monitored with a capacitive sensor attached to the metal lick spout. For some of the initial animals, no lick data were recorded. Mice alternately ran on the two tracks for a total of 15–30 runs on each track and day. The mice made 1–5 runs on one track and then an equal number of runs on the other. The length of these trial blocks was randomly varied. Imaging was performed with the same set of contexts for at least two (up to five) subsequent days. In many of the mice, the visible area under the imaging window was sufficiently large to select another imaging field of view that contained a different population of neurons. In these cases, we repeated the entire imaging experiment after the first experiment had been completed and used the new field of view and a different novel context (see Extended Data Fig. 5). In this manner, we performed a total of twelve experiments in six animals for the DG, seven experiments in five animals for CA3, and fifteen experiments in eleven animals for CA1 in the familiar–novel paradigm, as well as six experiments in three mice for DG imaging in the simplified environment, and seven experiments in six animals for the poor–normal–rich paradigm in CA1.

**In vivo two-photon calcium imaging.** Imaging was performed using a resonant/galvo high-speed laser scanning two-photon microscope (Neurolabware) with a frame rate of 30 Hz for bidirectional scanning. The microscope was equipped with an electrically tunable, fast  $z$ -focusing lens (optotune, Edmund Optics) to switch between  $z$ -planes within less than a millisecond. Images were acquired through a 16× objective (Nikon, 0.8 N.A., 3 mm WD), which was tilted at an angle of 20° for CA3 imaging. GCaMP6f was excited at 930 nm with a femtosecond-pulsed two-photon laser (Mai Tai DeepSee, Spectra-Physics). We scanned three imaging planes (~25 µm  $z$ -spacing between planes) in rapid alternation so that each plane was sampled at 10 Hz. The planes spanned 300–500 µm in the  $x/y$ -direction and were placed so that as many principal cells as possible were depicted. An early subset of the CA1 imaging experiments was performed with a galvo/galvo laser scanning microscope (Femto 2D, Femtonics) and a Chameleon Ultra II two-photon laser (Coherent) using only a single plane for imaging. To block ambient light from the photodetectors, the animal's head plate was attached to the bottom of an opaque imaging chamber before each experiment, and the chamber was fixed in the behavioural apparatus together with the animal. A ring of black foam rubber between the imaging chamber and the microscope objective blocked any remaining stray light.

**Histology and imaging area detection.** At the end of experiments, mice were deeply anaesthetized using ketamine/xylazine (Sigma Aldrich) and image stacks of the area underneath the imaging window were acquired *in vivo* in the two-photon microscope (see Supplementary Video 2). Mice were then perfused transcardially with 4% paraformaldehyde in PBS. Brains were cut into 100-µm coronal slices and sections containing the area underneath the imaging window were collected. Image stacks of GCaMP6f and tdT fluorescence in the sections were acquired with



a confocal microscope (LSM 710, Zeiss) and the imaged region was re-identified by comparing these stacks with the ones obtained *in vivo*, as described above (Fig. 1c, Extended Data Fig. 2c). For CA2/3 imaging experiments, the location of the imaged areas in CA3 and CA2 was confirmed for all mice by referencing sections to an anatomical atlas<sup>36</sup>.

**Imaging data processing, segmentation and data extraction.** Motion correction of all imaging data was performed line-by-line using the SIMA software package<sup>37</sup> with a 2D hidden Markov model<sup>38</sup>. When necessary, a pre-alignment step was performed using Matlab (version R2015b, MathWorks) built-in functionality (rigid transformation). Motion artefacts were estimated on either the red tdT or the green GCaMP6f fluorescence channel, whichever gave the better result for a given dataset. If no decent motion correction could be achieved, the data were discarded. Next, the motion-corrected and time-averaged image of tdT for each run was used to align recordings from the same field of view relative to each other and their displacements were stored with the dataset.

For obtaining data from principal cells, regions of interest (ROIs) were drawn manually around cell bodies (segmentation) in the principal cell layers of the three hippocampal subfields using ImageJ (NIH). Cell bodies were identified based on the motion-corrected, time-averaged GCaMP6f fluorescence images and re-inspected for each run to make sure that segmented cells were clearly visible throughout the experiment. We disambiguated GCs from other cell types by their small soma size and their location in the granule cell layer of the DG. We did not segment or include any neurons into this GC dataset that had unusually large somata or were located in the hilar region. Thereby, we made sure that our data represent the activity of GCs rather than hilar mossy cells or GABAergic DG interneurons<sup>39,40</sup>. In CA1 and CA2/3, only neurons in the pyramidal cell layers were segmented and included in the PYR dataset. Neurons in the pyramidal cell layer that co-expressed tdT were excluded from this dataset. For the disambiguation of superficial- and deep-layer CA1 PYRs (Extended Data Fig. 9), we referenced our imaging planes with the 3D stacks of the entire hippocampal formation from the respective mice and manually selected PYR somata in regions that were close ( $\sim 50\mu\text{m}$ ) to the oriens border for the deep and somata close to the radiatum border for the superficial group. PVI somata were identified in the red tdT channel and ROIs were drawn around their somata in stratum oriens, pyramidal and radiatum in CA1 and in the granule cell layer as well as the hilar region in the DG.

The obtained ROIs were transformed according to the displacement between the mean GCaMP6f fluorescence images as described above, and the average calcium signal over time was obtained from each ROIs for all runs. We restricted further analysis on running periods with a speed of at least  $5\text{ cm s}^{-1}$ . We identified significant calcium transients, which reflect the firing of principal cells, as described<sup>31,38</sup>. In brief, calcium traces were corrected for slow changes in fluorescence by subtracting the eighth percentile value of the fluorescence-value distribution in a window of  $\sim 8\text{ s}$  around each time point from the raw fluorescence trace. We obtained an initial estimate on baseline fluorescence and standard deviation (s.d.) by calculating the mean of all points of the fluorescence signal that did not exceed 3 s.d. of the total signal and would therefore be likely to be part of a significant transient. We divided the raw fluorescence trace by this value to obtain a  $\Delta F/F$  trace. We used this trace to determine the parameters for transient detection that yielded a false positive rate (defined as the ratio of negative to positive oriented transients)  $< 5\%$  and extracted all significant transients from the raw  $\Delta F/F$  trace. Definitive values for baseline fluorescence and baseline s.d. were then calculated from all points of this trace that did not contain significant transients. For further analysis, all values of this  $\Delta F/F$  trace that did not contain significant calcium transients were masked and set to zero.

For PVIs, rigorous identification of spike or burst-related calcium transients was prevented by their high firing rates, which exceeds the acquisition rate of our imaging system and the kinetics of the calcium indicator<sup>41–46</sup>. However, the calcium signal from PVIs can still be used to approximate PVI firing rate over time<sup>13,20,45,46</sup>. To obtain baseline-normalized  $\Delta F/F$  calcium traces from PVIs, we extracted their raw fluorescence signals and divided them by their mean fluorescence in periods in which the virtual environment screens were blanked between the runs. Therefore, the  $\Delta F/F$  trace of PVIs reports the change in activity rates of the cells due to exposure to one of the contexts.

**Activity differences and spatial information.** Activity difference scores in Fig. 2a were calculated for each cell using the following formula:  $(\text{activity}_{\text{familiar}} - \text{activity}_{\text{novel}})/(\text{activity}_{\text{familiar}} + \text{activity}_{\text{novel}})$ . To calculate a measure for spatial information (SI) content for principal cells, we adapted a traditional method of SI assessment<sup>47</sup> to calcium imaging data. To calculate SI, the average calcium activity (mean  $\Delta F/F$ ) was computed for each 10-cm-wide bin along the linear track and used as an approximation for the neurons' average firing rate in that location. SI was then calculated for each cell as  $\text{SI} = \sum_{i=1}^N \lambda_i \ln \frac{\lambda_i}{p_i}$  in which  $\lambda_i$  and  $p_i$  are the average calcium activity and fraction of time spent in the  $i$ -th bin, respectively,  $\lambda$  is the overall calcium activity averaged over the entire linear track and  $N$  is the number

of bins on the track (40 in our case). Therefore, the amount of spatial information is inferred from differences in the calcium activity and reported as bits per s.

**Place field identification.** Place fields were identified according to published methods<sup>31,35</sup>. In brief, the mean  $\Delta F/F$  was calculated from significant calcium transients for 80 position bins (each 5-cm wide) and this mean fluorescence over distance plot was then smoothed by averaging over the three points adjacent to each bin. Potential place fields were initially identified as contiguous regions of this  $\Delta F/F$  over distance plot in which all of the points were greater than 25% of the difference between the bin with the highest  $\Delta F/F$  value and the baseline value (mean of the lowest 20 out of 80 bins'  $\Delta F/F$  values). In addition, the candidate place fields had to fulfil the following criteria: (1) the potential field had to have a width of at least 3 bins (corresponds to 15 cm running distance on the ball circumference); (2) the mean  $\Delta F/F$  value inside the field had to be at least seven times the mean of the  $\Delta F/F$  value outside the field; and (3) significant calcium transients had to be present at least 20% (for CA1–3 PYRs) and 10% (for GCs) of the time in which the mouse was moving in the field. Potential place fields that fulfilled these criteria were accepted if their  $P$  value from bootstrapping exceeded 0.05. For bootstrapping, the  $\Delta F/F$  trace for each experiment was broken into segments of at least 50 consecutive imaging frames and randomly shuffled. This was repeated 1,000 times for each cell. Then the place field detection procedure, as described above, was performed on each of the shuffled  $\Delta F/F$  traces. The  $P$  value of the place field was then defined as the number of these randomly shuffled traces on which a place field was detected according to the outlined criteria divided by the number of shuffles (1,000). Overall, the criteria for place cell identification were relatively conservative and may underestimate the fraction of detected place cells among the active cells (Fig. 1f). The best parameters for place field detection were determined on the first obtained datasets by systematic variation and optimized to detect the maximum number of significant place fields.

**Place field stability, consistency and discrimination between contexts.** To assess the similarity of a place cell's spatial representation on different contexts or days, we first divided the track into 40 bins and calculated the mean  $\Delta F/F$  value for each bin on the track, based on all significant calcium transients (activity map) for each individual cell. One of these maps was computed for each context and day in all cells. The stability of place field of a cell was measured as the cross-correlation of the mean activity maps for runs in the same context on two different days. To assess whether the mean place field stability of cells in a region between a given day and a target day was above chance (Fig. 4d; Extended Data Fig. 10c), we generated chance distributions by correlating each cell's activity on the selected day with that of a randomly chosen cell on the target day. The consistency of place field firing was determined as the cross-correlation between the average activity of the first and the second block of five consecutive runs on the same track and session. Finally, the similarity between contexts was determined as the correlation of mean activity maps for runs in the familiar context and runs in the novel context on the same day (Figs. 2d, 3g, Extended Data Fig. 3g). To calculate the trial-to-trial reliability (Fig. 3f, Extended Data Figs. 3i, 7e), we calculated the pairwise cross-correlations between the calcium signals of all individual runs in one session on the same track and averaged the obtained values for each cell. We computed PoVs from the activity of all imaged cells in a session by stacking their average activities over distance on top of each other. PoV correlations between days or contexts were then determined as the cross-correlation of these cellular activities in each 10-cm bin along the linear track between two different contexts or days<sup>10,48</sup>.

**Place field shift and centre of mass.** To assess the shift of place fields (Extended Data Fig. 7c), we first calculated the centre of mass (COM) for each place field based on the mean activity map over track distance according to the following equation:  $\text{COM} = \frac{\sum_i \Delta F_i x_i}{\sum_i \Delta F_i}$  in which  $N$  is the number of bins covered by the place field,  $\Delta F_i$  is the fluorescence in the  $i$ -th bin of the place field and  $x_i$  is the distance of the  $i$ -th bin from the start of the track. For all cells that had a defined place field on the same linear track on two consecutive days, the place field relocation distance was defined as the distance between the two COMs of the place fields.

**Statistics.** All statistical tests are described in the corresponding figure legends. All comparisons were two-sided. Unless indicated otherwise, statistical comparisons were made between cells fulfilling the individual criteria as specified in the figure legend. The reported  $n$  numbers exclude missing ('NaN') values. All ANOVA tests are one-way tests.

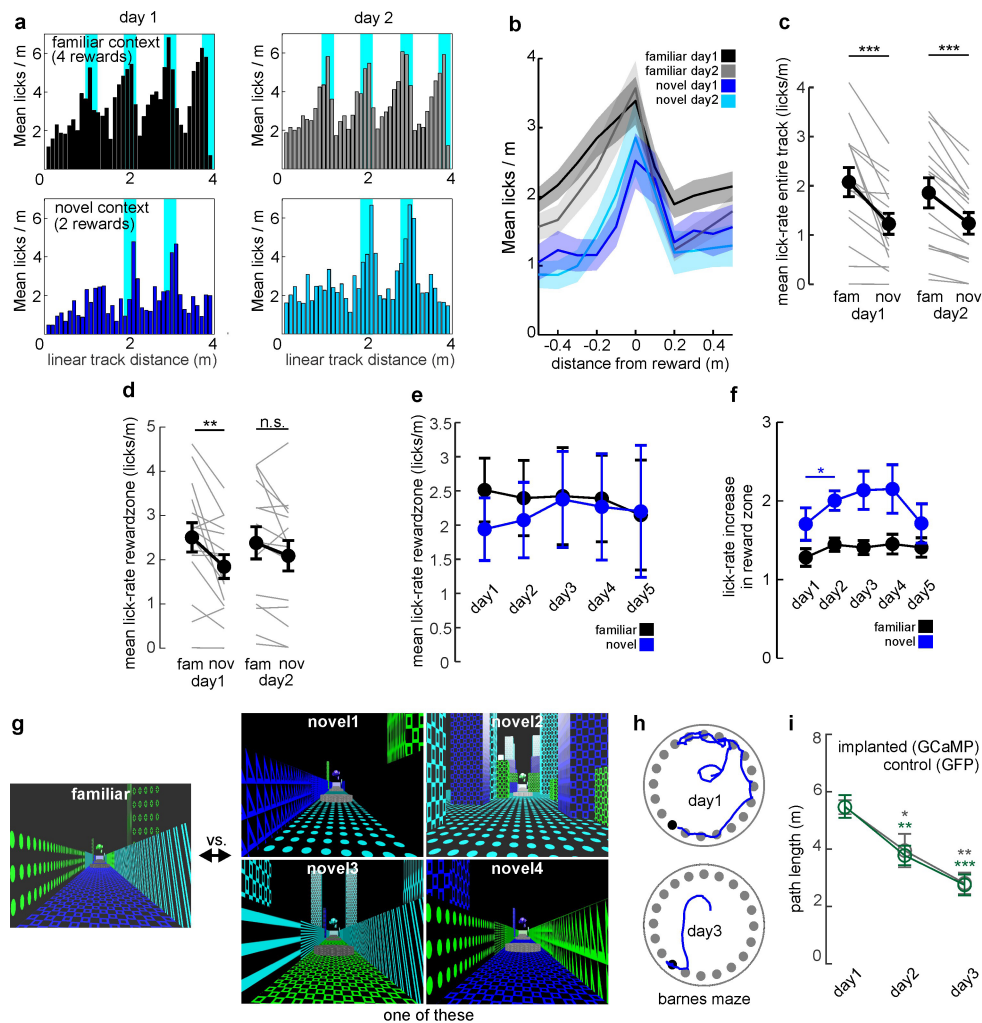
**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

**Code availability.** Any custom written code is available upon request.

**Data availability.** The data that support the findings of this study are available from the corresponding author upon reasonable request.

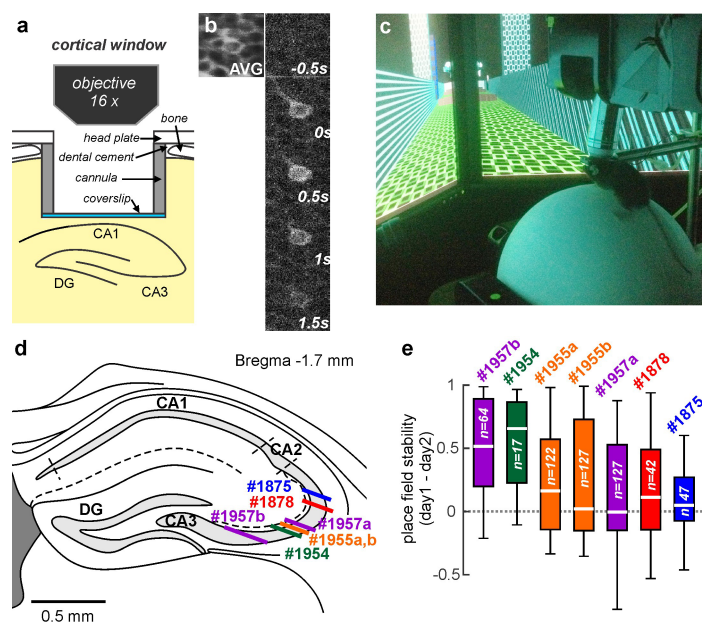
- Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L. & Tank, D. W. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat. Neurosci.* **13**, 1433–1440 (2010).

32. Malvache, A., Reichinnek, S., Villette, V., Haimerl, C. & Cossart, R. Awake hippocampal reactivations project onto orthogonal neuronal assemblies. *Science* **353**, 1280–1283 (2016).
33. Schmidt-Hieber, C. & Häusser, M. Cellular mechanisms of spatial navigation in the medial entorhinal cortex. *Nat. Neurosci.* **16**, 325–331 (2013).
34. Aghajian, Z. M. et al. Impaired spatial selectivity and intact phase precession in two-dimensional virtual reality. *Nat. Neurosci.* **18**, 121–128 (2015).
35. Sheffield, M. E. J. & Dombeck, D. A. Calcium transient prevalence across the dendritic arbour predicts place field properties. *Nature* **517**, 200–204 (2015).
36. Franklin, K. B. J. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates* (Elsevier, Amsterdam, 2008).
37. Kaifosh, P., Zaremba, J. D., Danielson, N. B. & Losonczy, A. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Front. Neuroinform.* **8**, 80 (2014).
38. Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L. & Tank, D. W. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56**, 43–57 (2007).
39. Hosp, J. A. et al. Morpho-physiological criteria divide dentate gyrus interneurons into classes. *Hippocampus* **24**, 189–203 (2014).
40. Hainmüller, T., Kriegelstein, K., Kulik, A. & Bartos, M. Joint CP-AMPA and group I mGlu receptor activation is required for synaptic plasticity in dentate gyrus fast-spiking interneurons. *Proc. Natl Acad. Sci. USA* **111**, 13211–13216 (2014).
41. Bartos, M., Vida, I. & Jonas, P. Synaptic mechanisms of synchronized gamma oscillations in inhibitory interneuron networks. *Nat. Rev. Neurosci.* **8**, 45–56 (2007).
42. Varga, C., Golshani, P. & Soltesz, I. Frequency-invariant temporal ordering of interneuronal discharges during hippocampal oscillations in awake mice. *Proc. Natl Acad. Sci. USA* **109**, E2726–E2734 (2012).
43. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
44. Tukker, J. J. et al. Distinct dendritic arborization and in vivo firing patterns of parvalbumin-expressing basket cells in the hippocampal area CA3. *J. Neurosci.* **33**, 6809–6825 (2013).
45. Arriaga, M. & Han, E. B. Dedicated hippocampal inhibitory networks for locomotion and immobility. *J. Neurosci.* **37**, 9222–9238 (2017).
46. Garcia-Junco-Clemente, P. et al. An inhibitory pull-push circuit in frontal cortex. *Nat. Neurosci.* **20**, 389–392 (2017).
47. Skaggs, W. E., McNaughton, B. L., Gothard, K. M. & Markus, E. J. An information-theoretic approach to deciphering the hippocampal code. In *Advances in Neural Information Processing Systems (NIPS)* 1030–1037 (1993).
48. Danielson, N. B. et al. Sublayer-specific coding dynamics during spatial navigation and learning in hippocampal area CA1. *Neuron* **91**, 652–665 (2016).



**Extended Data Fig. 1 | Virtual environment behavioural paradigm for head-fixed mice.** Related to Fig. 1. **a**, Mean number of licks per spatial bin (10-cm wide) for one example mouse on day 1 (left) and day 2 (right) on the familiar (top, grey) and novel (bottom, blue) linear tracks. Blue shaded areas indicate reward zones. **b**, Mean lick rate per bin as a function of distance from the next reward location for the familiar (black traces) and novel (blue traces) contexts. Shaded areas denote s.e.m.,  $n = 15$  experiments. **c**, Mean lick rate over the entire familiar (left) or novel (right) track on day 1 (left) or day 2 (right). Grey lines denote individual experiments ( $n = 15$ ) and black circles with error bars show mean  $\pm$  s.e.m. **d**, As in c but for mean lick rate in the reward zones only. **e**, Reward-related licking plotted for the experiments continued over 5 days. In this subset of the data ( $n = 5$  experiments) no significant difference in licking between contexts was observed on any day (repeated-measures one-way ANOVA), although there was a trend towards lower lick rates in the novel context. **f**, Lick rate increase in the reward zone (as compared to licking on the remaining track). An increase in the fraction of reward-related licks was

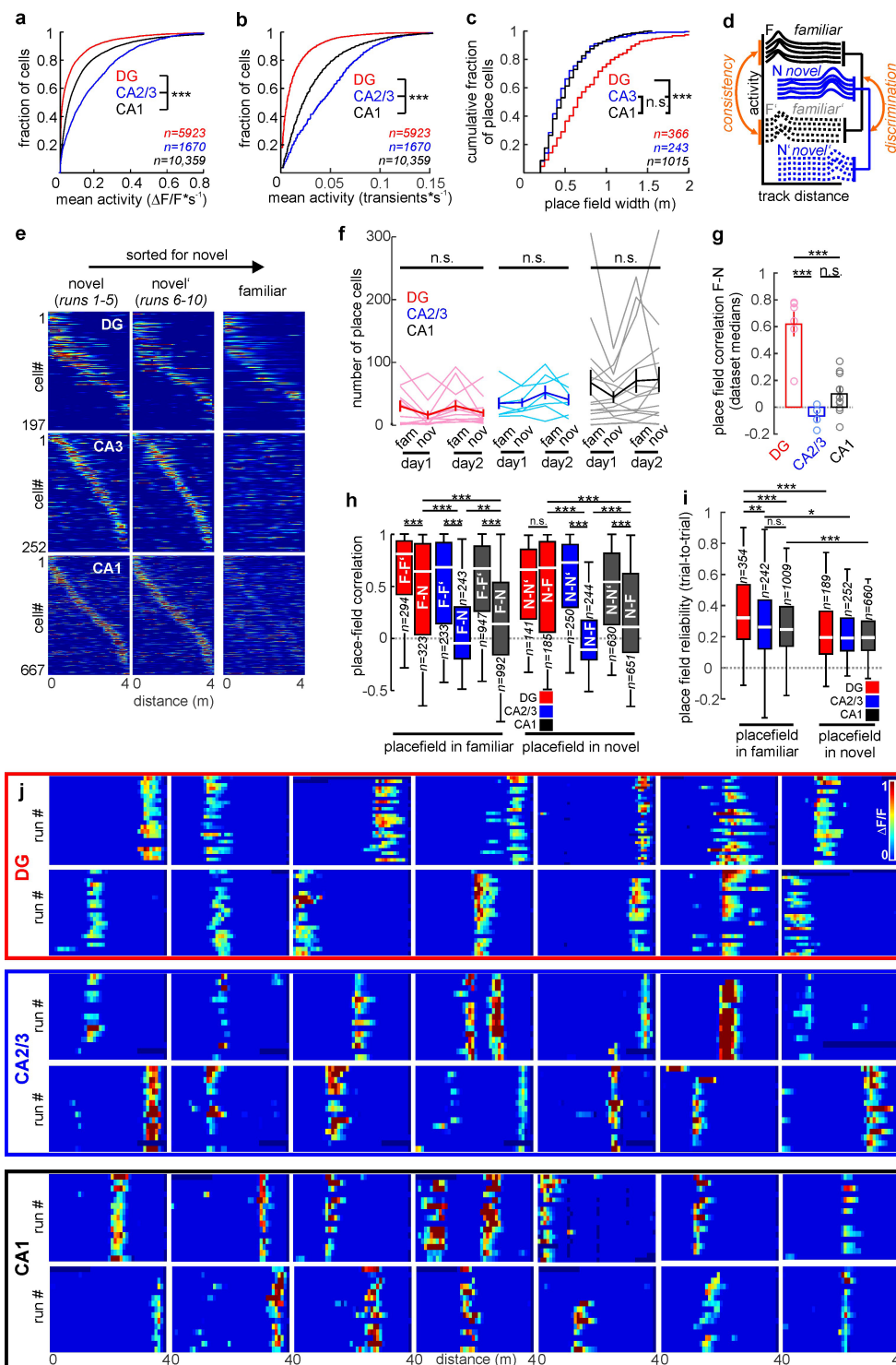
observed only between day 1 and 2 ( $n = 5$  experiments). **g**, Screenshots of the familiar context and the four different novel context sceneries. One novel context was randomly selected for each experiment from this set and maintained through all days. If more than one experiment was performed using a given animal, a different novel context was chosen for each of the experiments. **h**, Intact spatial memory was probed in the experimental mice and GFP-injected controls in a Barnes maze learning paradigm (see Methods). Blue traces show a mouse's trajectory in the last of four sessions on the first (top) and third (bottom) days of the experiment. **i**, Mean path length per session for implanted mice used for imaging experiments ( $n = 8$  mice; green) and GFP-injected controls ( $n = 8$  mice; grey) on days 1–3 (repeated-measures one-way ANOVA). There was no significant difference in path length between groups (day 1:  $P = 0.96$ , day 2:  $P = 0.806$ , day 3:  $P = 0.915$ , two-sided  $t$ -test). **c**, **d**, **f**, Two-sided paired  $t$ -test.  $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ , n.s. not significant. Error bars denote s.e.m. throughout. For exact  $P$  values see Supplementary Table 1.



**Extended Data Fig. 2 | Imaging configuration and CA3 imaging locations.** **a**, Schematic of the transcortical window implantation. A stainless steel cannula (3 mm diameter) with a circular coverslip attached to the bottom is implanted into the brain and rests on the external capsule on top of the hippocampus (see Methods). **b**, Illustrative fluorescence time series of an active GC, including the time-averaged GCaMP fluorescence (AVG, left). **c**, Photograph of a mouse in the virtual environment setup. Depicted is a mouse in the CA2/3 recording group with tilted objective for lateral access view (see Methods). **d**, Anatomical drawing of the dorsal hippocampus indicating the imaging planes for CA3 recordings.

Depicted is the location of the middle plane from a three-plane imaging volume of 40–80  $\mu\text{m}$ . Coloured numbers indicate individual mice from which the data are derived; in cases for which more than one experiment was performed in a mouse, letters denote the locations of the respective experiments. **e**, Box plot of place field stability values over days from familiar-context place cells in the individual recordings shown in **d**. White numbers denote the number of place cells from each experiment that are represented by the boxes. Boxes, 25th to 75th percentiles; white bars, median; whiskers, 99% range.

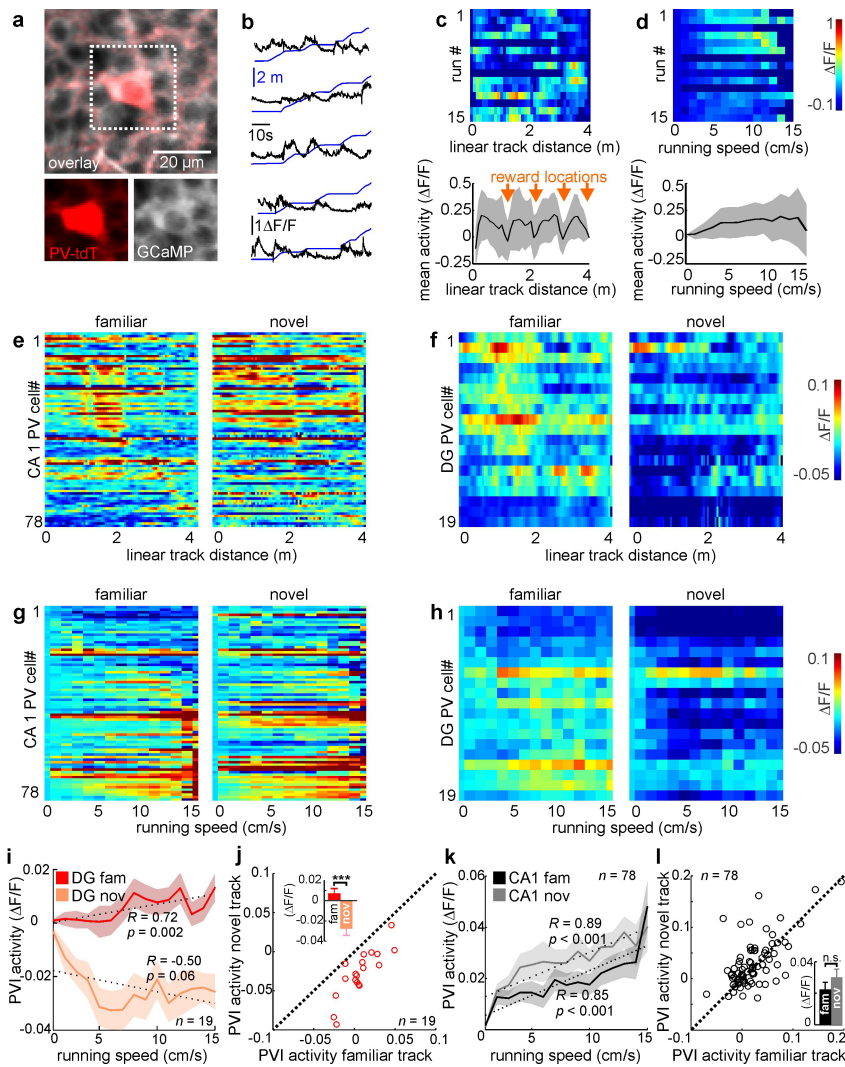




Extended Data Fig. 3 | See next page for caption.

**Extended Data Fig. 3 | Properties of place coding in the familiar and novel virtual contexts.** Related to Fig. 2. **a**, Cumulative distribution of mean calcium activity, based on the area under the calcium curve, for all GCs (red), CA2/3 PYRs (blue) and CA1 PYRs (black) recorded in the familiar context. **b**, As in **a**, but for the calcium transient rate. **c**, Cumulative distribution of place field widths for all day 1 familiar-context place cells in the respective hippocampal areas. **d**, Experimental schematic: place field consistency was measured as the correlation of the average activity on the first and second blocks of five runs in the same context (familiar: F–F' or novel: N–N'). Place field discrimination was assessed as the correlation of activity maps for different contexts (F–N). **e**, Activity over distance for cells with a place field on the novel track. Cells were sorted according to their peak activity in novel-track runs and are plotted separately for the first (left) and second (middle) blocks of runs on the novel track and for all runs on the familiar track (right). For the same plots with familiar-context place cells, see Fig. 2c. **f**, Number of cells with significant place fields (see Methods) in each experiment. Thin lines denote individual experiments (DG,  $n = 12$ ; CA2/3,  $n = 7$ ; CA1,  $n = 15$  experiments), thick lines with error bars show mean  $\pm$  s.e.m. (one-way repeated measures ANOVA). **g**, Median activity-map

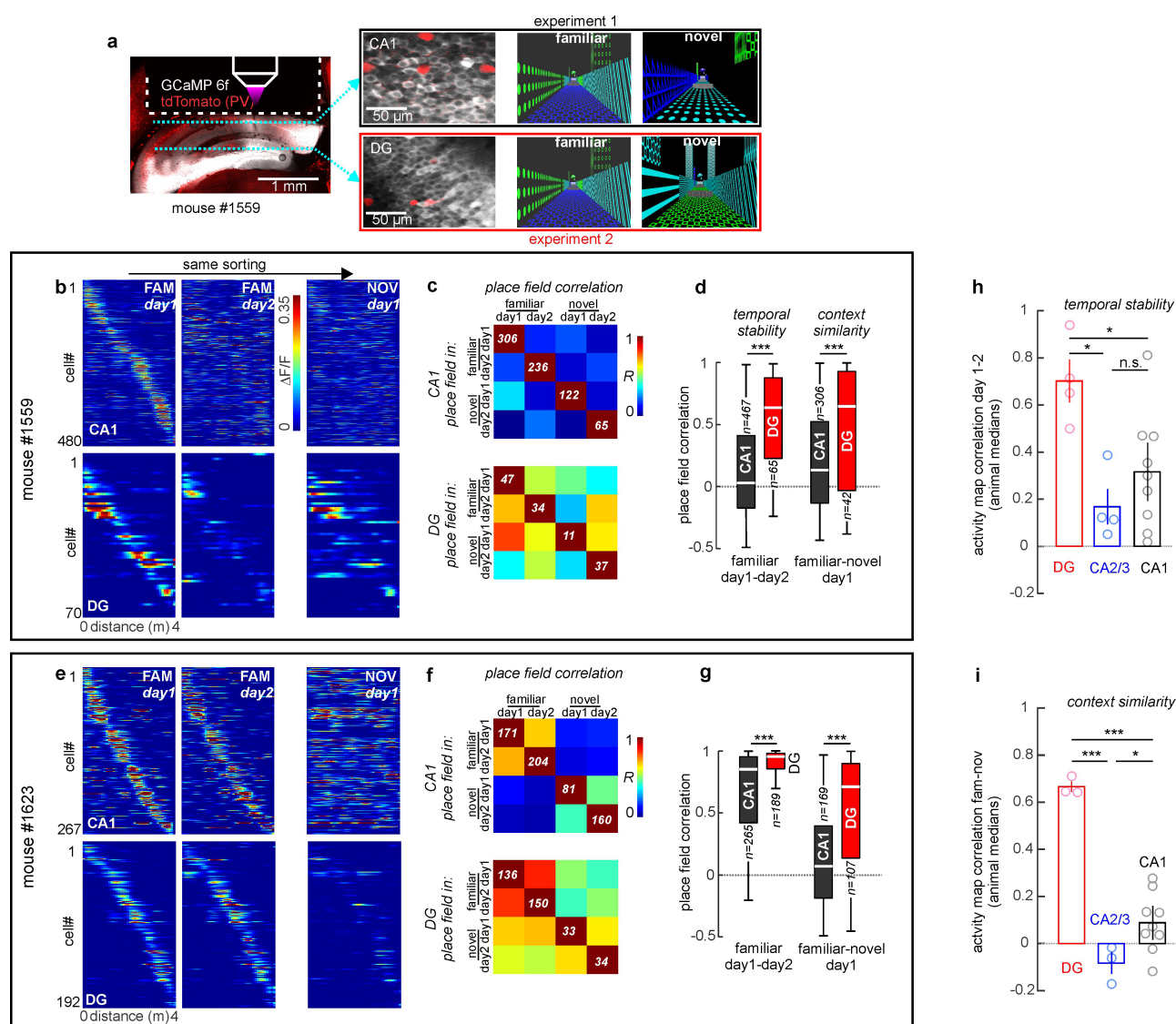
correlations between the familiar and novel contexts for experiments with at least 20 place cells on day 1 (circles; DG,  $n = 6$ ; CA2/3,  $n = 4$ ; CA1,  $n = 11$  experiments) and their mean  $\pm$  s.e.m. (ANOVA, Holm–Sidak). **h**, Pearson's correlation of place-related activity for cells that had a place field in the familiar context (left) or novel context (right) on day 1. Left box of each pair shows correlation of activity maps for trial blocks on the same track ('consistency') and right bar shows correlations with trials on the other track ('discrimination') in the same session. Correlations within the same context were always significantly higher than between contexts, except for GCs that had newly acquired a place field on the novel track. **i**, Trial-to-trial reliability of place cell responses. Place-related firing was more reliable in the familiar than in the novel context and higher in GCs than in PYRs in the familiar context. Differences between areas were not significant for the novel context. **j**, Each figure shows the run-by-run calcium activity (colour coded) over distance from one individual place cell for one session in the familiar context. Rows denote individual runs. **h**, **i**, Boxes, 25th to 75th percentiles; white bars, median; whiskers, 99% range. **a–c**, **h**, **i**, One-way ANOVA on ranks, Dunn's test. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ; n.s., not significant. For exact  $P$  values see Supplementary Table 1.



#### Extended Data Fig. 4 | Task-related activity of PV-expressing

**interneurons in CA1 and DG.** **a**, Illustrative, time-averaged fluorescence image of GCaMP6f (pseudocolour white) and tdT (red) expressed in PVIs. Insets below show each fluorescence channel separately for the area indicated with the dotted line. **b**, Calcium trace of a representative PVI in CA1 (black) and distance on the virtual linear track (blue) over time. This PVI is active particularly at times when the animal moves fast. **c**, Calcium activity (colour coded) of the representative cell shown in **b** as a function of linear track distance for multiple runs (rows). Graph at the bottom shows the average of activity over distance for this cell in the familiar context. Shaded areas denote s.d. The same analysis was performed for 78 CA1 and 19 DG PVIs. **d**, As in **c**, but activity was plotted as a function

of running speed. **e**, **f**, Mean activity maps over distance for all recorded CA1 (**e**) or DG (**f**) PVIs in the familiar and novel contexts. **g**, **h**, Same as in **e**, **f**, but mean activity was plotted over running speed. Activity in most DG PVIs is suppressed in the novel context. **i**, Activity of PVIs in the DG on the familiar (red) and novel track (light red) over running speed. Shaded areas denote s.d.  $R$  denotes Pearson's  $R$ . **j**, Mean calcium activity during familiar-track running plotted against novel-track activity for PVIs in the DG. Inset, bars denote mean  $\pm$  s.e.m. of running-related activity for novel and familiar tracks (two-sided signed rank-sum test). **k**, **l**, As in **i**, **j** but for CA1 PVIs. \*\*\* $P < 0.001$ ; n.s., not significant. For exact  $P$  values see Supplementary Table 1.

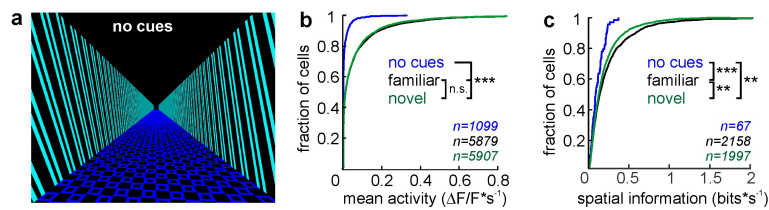


**Extended Data Fig. 5 | Differential stability and context discrimination of CA1 and DG are not due to interindividual differences.**

**a**, Experimental schematic. Left, fluorescence image of GCaMP6f (white) and tdT in PVIs (red) in a post mortem coronal brain section. Dotted line indicates position of the imaging window. In this mouse, recordings were made from CA1 PYRs and DG GCs in separate, sequential experiments (exemplary image planes are shown in middle images). For each experiment, the same familiar and a different novel context (right) were used. Thereby, the coding properties of PYRs and GCs in the same animal could be compared. **b**, Calcium activity over distance for CA1 PYRs (top) and DG GCs (bottom) with place fields on the familiar track sorted for their peak activity on day 1 in the familiar context. Activity of the same cells with the same sorting on day 2 in the familiar context (middle) and on day 1 in the novel context (right). **c**, Mean cellular activity map correlations (colour coded: Pearson's  $R$ ) over two days and contexts as indicated on the  $x$ -axis. Data sampled only for place cells recorded in the selected mouse. Each row shows mean correlation values for cells that had a place field on the day and track indicated on the  $y$ -axis ( $n$  denoted as white numbers). **d**, Left, correlations of activity maps in the familiar context between days plotted for cells that had a place field in the familiar

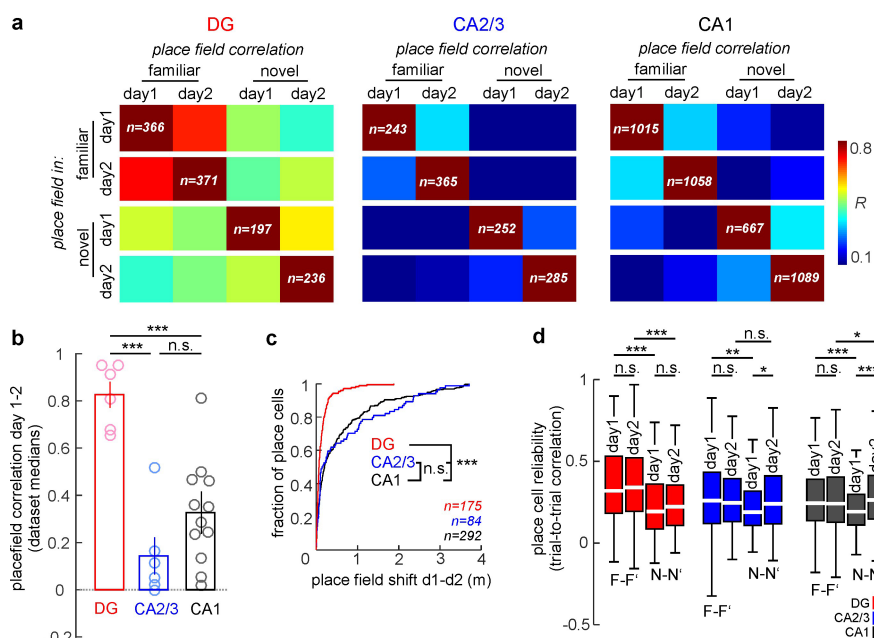
context. Cells were sampled only from measurements in this particular mouse. Right, activity map correlations between the familiar and novel contexts on day 1 for all cells that had a place field in the familiar context on that day. Stability over time and activity map similarity between contexts are significantly higher for GCs than for CA1 PYRs in the same mouse. **e–g**, As in **b–d**, but calculated separately on cells from another mouse. **h**, Activity map correlations between days 1 and 2 were calculated for familiar-context place cells and medians (circles) are displayed for each animal that had a minimum of 20 such place cells (DG,  $n = 4$ ; CA2/3,  $n = 4$ ; CA1,  $n = 9$  mice). The means  $\pm$  s.e.m. of these per-animal medians (bars) were compared statistically. **i**, As in **h**, but for activity map correlations of familiar-context place cells on day 1 between the familiar and novel contexts (DG,  $n = 3$ ; CA2/3,  $n = 3$ ; CA1,  $n = 9$  mice). Higher temporal stability and higher inter-context similarity are a feature of GCs that is consistently observed in different mice. **h**, **i**, One-way ANOVA with Holm-Sidak test. Error bars denote s.e.m. **d**, **g**, Boxes, 25th to 75th percentiles; white bars, median; whiskers, 99% range. Two-sided rank-sum test. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ; n.s., not significant. For exact  $P$  values see Supplementary Table 1.





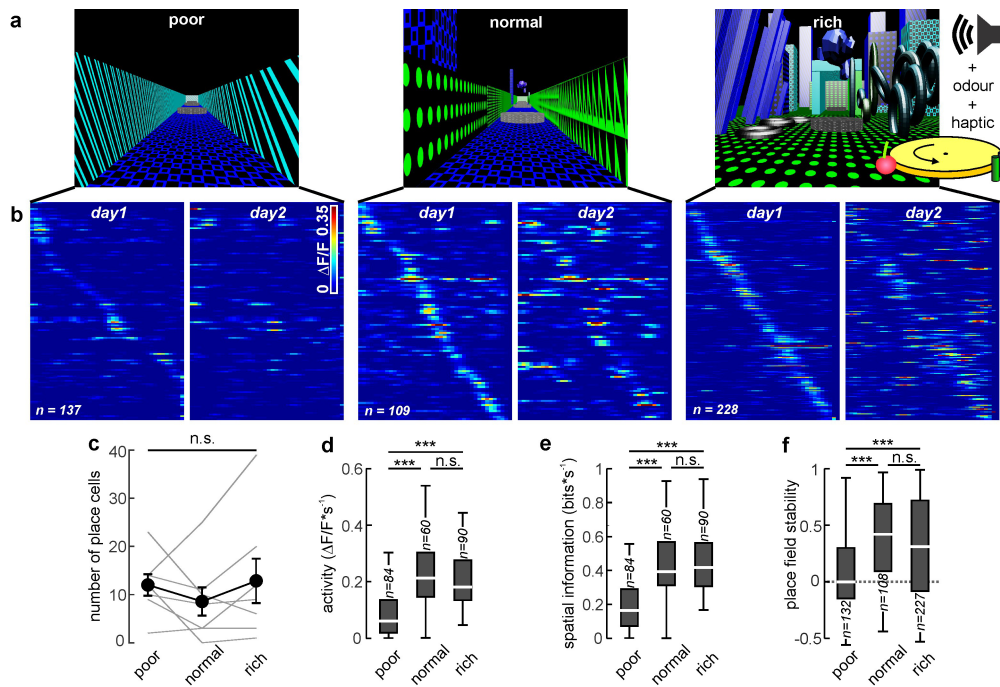
**Extended Data Fig. 6 | Spatial firing of DG GCs requires external reference cues.** **a**, Screenshot from the simplified virtual linear track devoid of any visual reference cues, except for patterned walls to provide a visual percept of self-motion. **b**, Cumulative distribution of activity levels for all GCs imaged in the standard familiar and novel contexts, as well as

the simplified version shown in **a**. **c**, Cumulative distribution of spatial information values in all cells with a minimum activity of 0.01 transients per second in the familiar, novel and self-motion-only based paradigms. \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ; n.s., not significant (one-way ANOVA on ranks with Dunn's test). For exact  $P$  values see Supplementary Table 1.



**Extended Data Fig. 7 | Place field remapping between days.** Related to Fig. 3. **a**, Mean cellular activity map correlations (Pearson's  $R$ ) over two days and contexts as indicated on the x-axis. Each row shows mean correlation values for cells (white numbers denote  $n$ ) that had a place field on the day and track indicated on the y-axis. **b**, Median day-to-day correlation ('stability') of familiar-context place cell activity for all experiments in which at least 20 cells had a place field in the familiar context on either day (circles; DG,  $n = 6$ ; CA2/3,  $n = 6$ ; CA1,  $n = 12$  experiments). Bars denote mean  $\pm$  s.e.m. of these per-experiment values (one-way ANOVA with Holm-Sidak test). **c**, For all cells that had a place field in the familiar context on both days of the experiment, the centres of mass for the activity in these place fields was determined. The graph

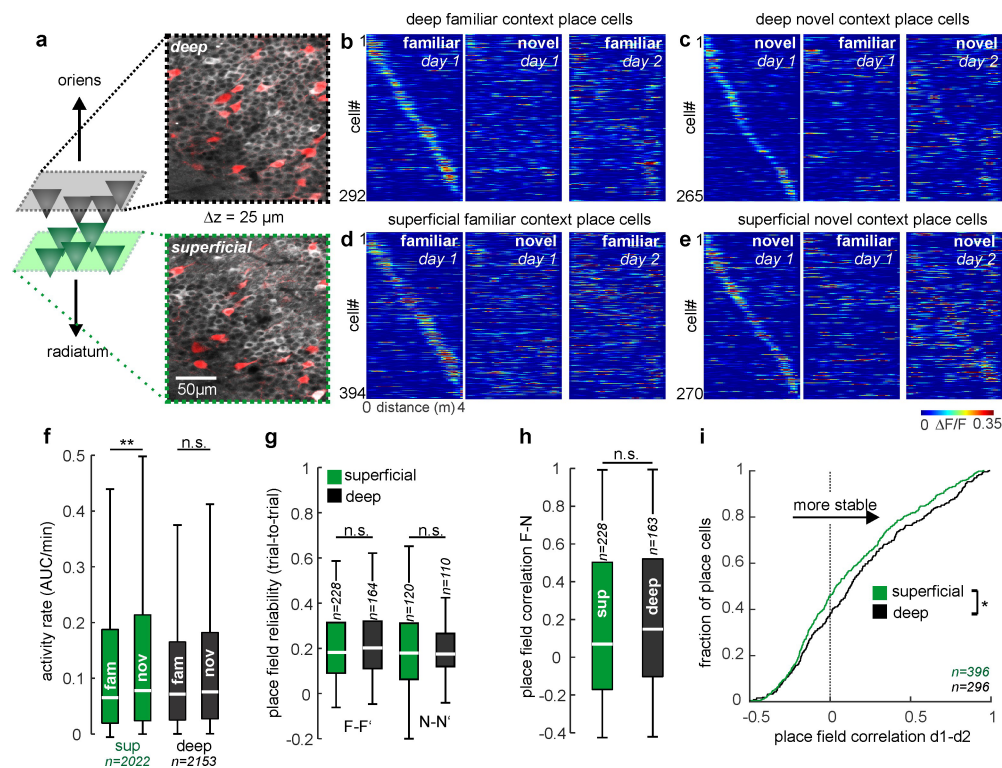
shows the cumulative distribution of the distances between these centres (shift) between days 1 and 2, which gives a measure of the relocation of place fields between days. **d**, Trial-by-trial correlation of place cell activity (reliability) in the familiar and novel contexts on days 1 and 2.  $N = 354$ , 367, 189, 226, 242, 364, 252, 285, 1,009, 1,044, 660 and 1,074 place cells per group (left to right). Place cell reliability for the novel context place cells increases in CA1 and CA2/3 between the first and second days. Boxes, 25th to 75th percentiles; white bars, median; whiskers, 99% range. **c, d**, One-way ANOVA on ranks with Dunn's test. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ; n.s., not significant. For exact  $P$  values see Supplementary Table 1.



**Extended Data Fig. 8 | CA1 place coding is degraded in the absence of visual references, but does not scale with environmental complexity.**

**a**, Screenshots of the three different virtual linear tracks. Left, the 'poor' track had patterned walls, but no other cues; middle, the 'normal' track with visual references; right, the 'rich' multisensory environment with many visual objects, sound, odour and tactile cues (Supplementary Video 6). **b**, Calcium activity over distance for CA1 PYRs with place fields on the tracks depicted above, sorted for their peak activity on day 1 (right) and, with the same sorting, on day 2 (left). Higher activity levels and day-to-day stability can be observed in the 'normal' and 'rich' environments.

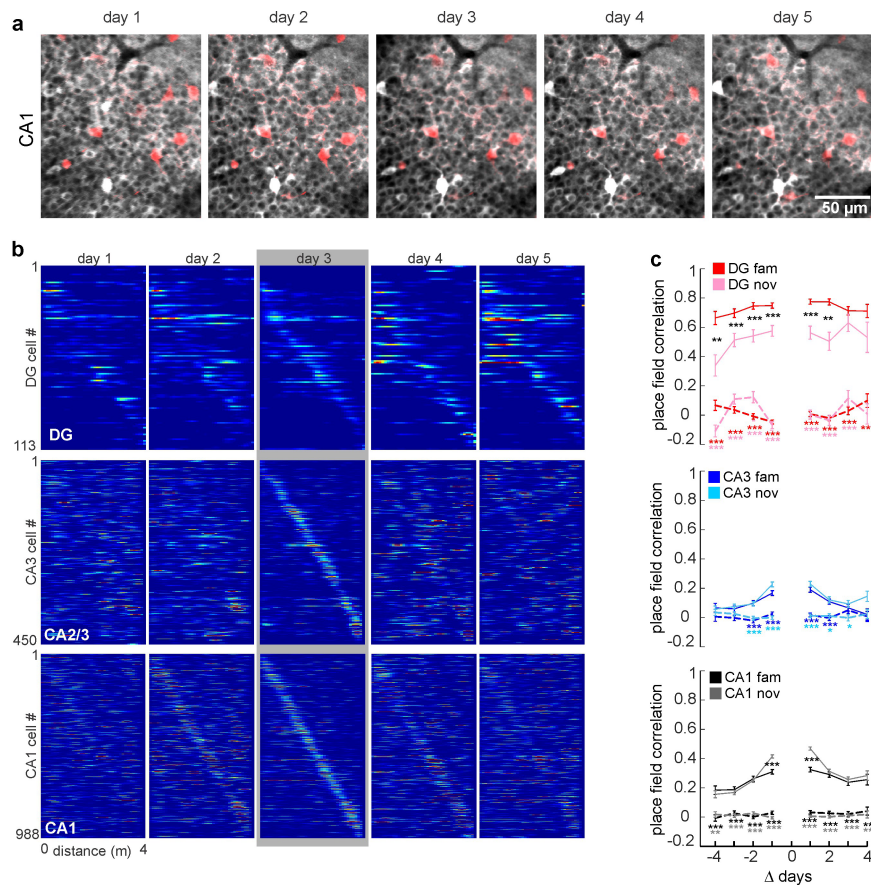
**c**, Number of cells with significant place fields (see Methods) on the first recording day per experiment. Thin lines denote individual experiments ( $n = 7$ ), thick lines with error bars the means  $\pm$  s.e.m. (repeated measures one-way ANOVA). **d**, Calcium activity levels (AUC) of the place cells detected in the three settings. **e**, As in **d**, but for spatial information. **f**, Activity map correlations between days for all cells that had a place field on the corresponding track. **d–f**, One-way ANOVA on ranks with Dunn's test. Boxes, 25th to 75th percentiles; white bars, median; whiskers, 99% range. \*\*\* $P < 0.001$ ; n.s., not significant. For exact  $P$  values see Supplementary Table 1.



**Extended Data Fig. 9 | Superficial and deep layer CA1 pyramidal cells differ in their task-related coding properties.** **a**, Experimental schematic. Cells with somata close to the border of the stratum oriens (deep CA1 PYRs) and those close to the border of the stratum radiatum (superficial CA1 PYRs) were identified in different z-planes and separated for analysis. Illustrative pictures to the right show time-averaged fluorescence of GCaMP6f (white) and td-Tomato (red) in PVIs. **b**, Calcium activity over distance for deep CA1 PYRs with a place field on the familiar track. Cells were sorted according to their peak activity on the familiar track (left) and are plotted in the same order for runs on day 1 on the novel track (middle) and for the runs on the familiar track on the second day (right). **c**, As in **b**, but for novel-context place cells on the familiar context (middle) and

novel context on the second day (right). **d**, **e**, As in **b**, **c**, but for superficial PYRs. **f**, Calcium activity rate for superficial and deep CA1 PYRs in the familiar and novel contexts. Activity rates increased significantly in the novel context in superficial but not in deep PYRs (two-sided signed rank-sum test). **g**, Trial-to-trial correlations of place cell activity compared between layers for familiar-context (left) and novel-context (right) runs. **h**, Activity map correlations between contexts for cells with a place field in the familiar context on day 1. **i**, Cumulative distribution of place cell activity map correlations between days 1 and 2 during familiar context runs. **g–i**, Two-sided rank-sum test. \* $P < 0.05$ , \*\* $P < 0.01$ ; n.s., not significant. Boxes, 25th to 75th percentiles; white bars, median; whiskers, 99% range. For exact  $P$  values see Supplementary Table 1.





**Extended Data Fig. 10 | Differential stability of hippocampal place fields over extended time spans.** Related to Fig. 4. **a**, The same place cells were imaged over multiple days. Pictures show an illustrative example of the time-averaged fluorescence of GCaMP6f (pseudocolour white) expressed pan-neuronally and tdT (red) expressed in PVIs for the same field of view in CA1 on five subsequent days. **b**, Activity maps for cells that had a place field on the novel track on any of the five days. Cells are sorted by their activity peaks on day three (grey shading). **c**, Activity map correlations as function of days passed for cells with place field on the familiar (dark colours) or novel track (light colours). Dotted lines

show corresponding levels of random correlations generated by shuffling cell IDs. Dark and light coloured asterisks underneath the dotted lines indicate significant differences of the actual versus random correlations for familiar- and novel-context place cells, respectively. Black asterisks between traces indicate significant differences between the mean correlation values of novel- and familiar-context place cells.  $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ . Two-sided rank-sum test for each time-difference (days) with Bonferroni correction. Error bars denote s.e.m. For exact  $P$  values and  $N$  numbers in **c** see Supplementary Table 1.

# A molecular rheostat adjusts auxin flux to promote root protophloem differentiation

P. Marhava<sup>1,5</sup>, A. E. L. Bassukas<sup>2,5</sup>, M. Zourelidou<sup>2</sup>, M. Kolb<sup>2,3</sup>, B. Moret<sup>1</sup>, A. Fastner<sup>3</sup>, W. X. Schulze<sup>4</sup>, P. Cattaneo<sup>1</sup>, U. Z. Hammes<sup>2,3</sup>, C. Schwechheimer<sup>2\*</sup> & C. S. Hardtke<sup>1\*</sup>

**Auxin influences plant development through several distinct concentration-dependent effects<sup>1</sup>. In the *Arabidopsis* root tip, polar auxin transport by PIN-FORMED (PIN) proteins creates a local auxin accumulation that is required for the maintenance of the stem-cell niche<sup>2–4</sup>. Proximally, stem-cell daughter cells divide repeatedly before they eventually differentiate. This developmental gradient is accompanied by a gradual decrease in auxin levels as cells divide, and subsequently by a gradual increase as the cells differentiate<sup>5,6</sup>. However, the timing of differentiation is not uniform across cell files. For instance, developing protophloem sieve elements (PPSEs) differentiate as neighbouring cells still divide. Here we show that PPSE differentiation involves local steepening of the post-meristematic auxin gradient. BREVIS RADIX (BRX) and PROTEIN KINASE ASSOCIATED WITH BRX (PAX) are interacting plasma-membrane-associated, polarly localized proteins that co-localize with PIN proteins at the rootward end of developing PPSEs. Both *brx* and *pax* mutants display impaired PPSE differentiation. Similar to other AGC-family kinases, PAX activates PIN-mediated auxin efflux, whereas BRX strongly dampens this stimulation. Efficient BRX plasma-membrane localization depends on PAX, but auxin negatively regulates BRX plasma-membrane association and promotes PAX activity. Thus, our data support a model in which BRX and PAX are elements of a molecular rheostat that modulates auxin flux through developing PPSEs, thereby timing PPSE differentiation.**

Auxin is a concentration-dependent permissive–restrictive signal in plant cell proliferation and differentiation–elongation that directly impinges on adaptive processes and growth rates<sup>1,2</sup>. Local auxin accumulations are important cues for organ organization. For example, high auxin concentration specifies the stem-cell niche in the *Arabidopsis* root tip<sup>2–4</sup>. Proximally, auxin concentration decreases gradually as stem-cell daughters repeatedly divide before they eventually differentiate. Notably, differentiation is accompanied by a renewed rise in auxin levels<sup>5,6</sup>. The underlying auxin distribution is generated by plasma-membrane-integral PINs, which are auxin efflux carriers with a coordinated asymmetric cellular localization that gives rise to directional polar auxin transport<sup>2–4</sup>. In root vasculature, PINs generally localize to the rootward end of cells, transporting auxin towards the root tip<sup>3</sup>. PINs are regulated by auxin, predominantly post-translationally<sup>7–9</sup>. Moreover, the AGC-family kinases D6 PROTEIN KINASE (D6PK) and PINOID (PID) activate auxin efflux through PIN phosphorylation<sup>10–12</sup>.

The proximo-distal auxin profile in root meristems intersects with differential auxin activity in the radial dimension. For example, developing PPSEs (Extended Data Fig. 1a, b) display higher auxin accumulation than surrounding cells<sup>6</sup> (Extended Data Fig. 2a–d) and differentiate, whereas neighbouring cells still remain meristematic<sup>6,13,14</sup> (Extended Data Fig. 2e). To explore whether PPSE differentiation depends on auxin activity, we manipulated the auxin response by expressing a constitutively active variant of an auxin-response factor, MONOPTEROS (MP<sup>A</sup>)<sup>15</sup>, under the control of PPSE-specific

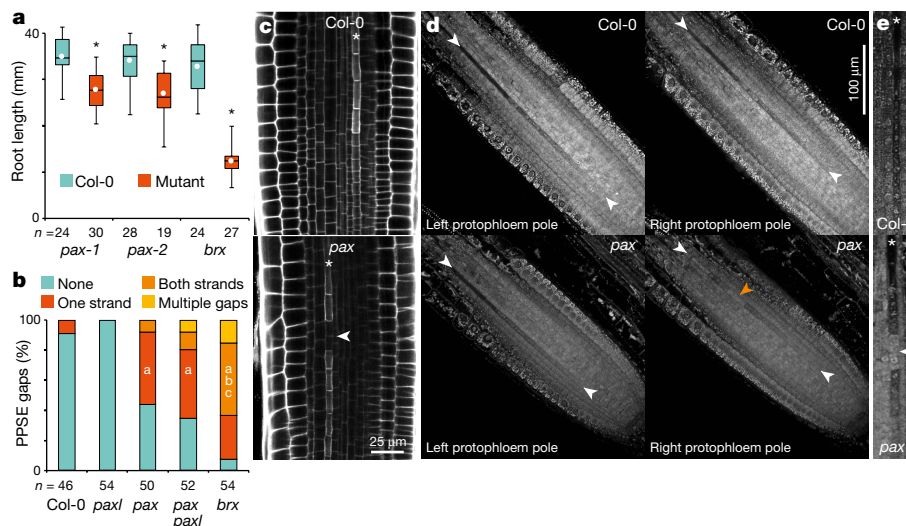
*COTYLEDON VASCULAR PATTERN 2 (CVP2)* promoter<sup>13,16</sup>. *CVP2::MP<sup>A</sup>* accelerated PPSE differentiation, indicating that auxin responses critically determine the differentiation process (Extended Data Fig. 2f, g).

How differential auxin activity is achieved in PPSEs remained unclear. BRX is plasma-membrane-associated, polarly localized and specifically expressed in developing PPSEs<sup>13,17</sup>. In *brx* mutants, PPSEs frequently fail to differentiate<sup>13</sup>. These cells lack the characteristic cell-wall changes and appear as gaps in the PPSE differentiation zone<sup>13,16,18</sup> (Extended Data Fig. 2h). A similar phenotype is observed in *octopus (ops)* mutants<sup>13,17</sup>, which are affected in a parallel genetic pathway required for PPSE differentiation<sup>19</sup>. Whereas OPS localizes to the shootward end of PPSEs, BRX co-localizes with PINs at the rootward end<sup>2,6,13,18</sup>. Auxin negatively regulates BRX protein abundance and plasma-membrane association, but induces *BRX* transcription<sup>18,20</sup>. Thus, BRX is a candidate for mediating auxin effects in PPSE differentiation. In *brx* PPSEs, auxin accumulation as compared to neighbouring cells was markedly lower and more variable than in the wild type (Extended Data Fig. 2i, j). Although *CVP2::MP<sup>A</sup>* expression in *brx* did not reduce the proportion of PPSE strands with gaps (Extended Data Fig. 2k), it significantly stimulated root growth (Extended Data Fig. 2l) and reduced gap size (Extended Data Fig. 2m). Such partial rescue was not observed with another PPSE-specific promoter that was inactive in gap cells (Extended Data Fig. 2k–n). Moreover, impaired PPSE differentiation was observed after pharmacological inhibition of auxin biosynthesis (Extended Data Fig. 2o, p), and *brx* protophloem defects were aggravated by genetic interference with auxin uptake (Extended Data Fig. 2q). These observations support the hypothesis that finely tuned auxin activity contributes to PPSE differentiation.

BRX protein is expressed only at low levels and in few cells, complicating cell-biological and biochemical investigations of BRX in its native context. However, a recently established trans-differentiation assay for sieve element formation<sup>21</sup> (Extended Data Fig. 3a, b) enabled us to perform proteomics analyses in a native cell type and identify specific BRX interactors by immunoprecipitation (Extended Data Fig. 3c, d). Among them, we retrieved D6PK and several D6PK-LIKE (D6PKL) kinases as well as PINs, but by far the most abundant was a D6PK/D6PKL-related kinase (AT2G44830)<sup>22</sup>, which we named PROTEIN KINASE ASSOCIATED WITH BRX (PAX).

To examine a potential role of AGC kinases in PPSE differentiation, we analysed *d6pk/d6pkl* as well as *pax* mutants. *D6PK/D6PKL* genes display substantial genetic redundancy and, consistent with normal PIN phosphorylation in their roots<sup>10,11,23</sup>, *d6pk0123* quadruple mutants had only a mild, possibly enhanced root-growth phenotype (Extended Data Fig. 4a). By contrast, *pax* loss-of-function mutants displayed reduced primary root growth (Fig. 1a), which was accompanied by PPSE differentiation defects (Fig. 1b–e). No phenotype was observed in a mutant of the closest PAX homologue, the uncharacterized PAX-LIKE (PAXL) kinase (AT5G40030)<sup>22</sup>, and *paxl* mutation only mildly enhanced the *pax* phenotype (Fig. 1b). A PAX–CITRINE fusion protein expressed

<sup>1</sup>Department of Plant Molecular Biology, University of Lausanne, Lausanne, Switzerland. <sup>2</sup>Plant Systems Biology, Technical University of Munich, Freising, Germany. <sup>3</sup>Department of Cell Biology and Plant Biochemistry, Regensburg University, Regensburg, Germany. <sup>4</sup>Department of Plant Systems Biology, University of Hohenheim, Stuttgart, Germany. <sup>5</sup>These authors contributed equally: P. Marhava, A. E. L. Bassukas. \*e-mail: claus.schwechheimer@wzw.tum.de; christian.hardtke@unil.ch



**Fig. 1 | Phenotypic characterization of *pax* mutants.** **a**, Root length of seven-day-old mutant and wild-type *A. thaliana* L. Heynh reference accession Columbia-0 (Col-0) seedlings. *pax-1* and *pax-2* are two independent PAX loss-of-function alleles. All data that are displayed subsequently were generated using *pax-1*. Box plots throughout show the second and third quartiles, maximum, minimum and mean (white dot). Statistically significant differences are indicated (two-sided Student's *t*-test; \**P* < 0.0008). **b**, Quantification of protophloem strands with gap cells, 6-day-old seedlings. Statistically significant differences are indicated (Fisher's exact test, two-sided, all *P* values < 0.001; a, significantly different

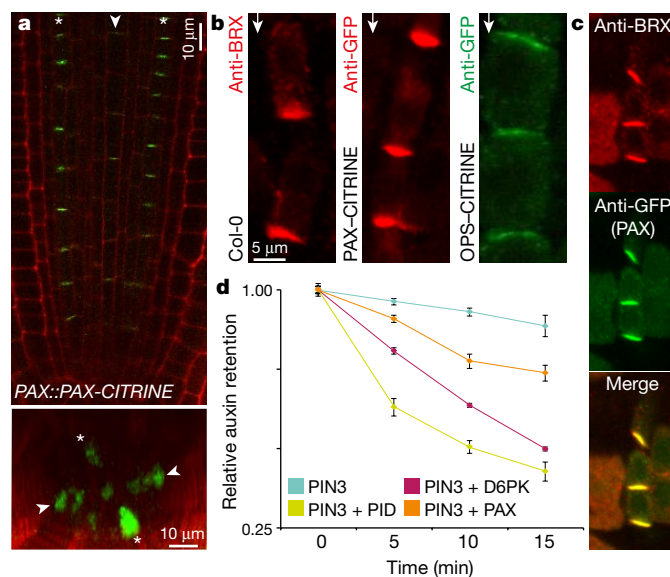
to Col-0; b, significantly different to *pax*; c, significantly different to *pax pax1*). **c**, Confocal microscopy of propidium iodide (PI)-stained root meristems. Asterisks indicate PPSE strands and the arrowhead indicates a gap cell in the *pax* protophloem. **d**, Confocal microscopy of Col-0 and *pax* root meristems (ClearSee fixation with PI staining), showing both protophloem poles. Note PPSE cell files (white arrowheads) that start with meristematic cells and end with mature empty sieve elements. In one *pax* pole, PPSE differentiation is perturbed (gap cells, orange arrowhead). **e**, Expanded view, highlighting a gap cell in a *pax* PPSE cell file (arrowhead). *n*, number of independent biological replicates.

under its native promoter complemented the PPSE differentiation phenotype of *pax* mutants (Extended Data Fig. 4b) and revealed PAX expression in developing protophloem, as well as weaker expression in the xylem axis (Fig. 2a). PAX displayed rootward cellular polarity (Fig. 2b) and co-localized with BRX (Fig. 2c). Exclusive expression of PAX-CITRINE in developing PPSEs, under the BRX promoter, fully rescued the *pax* protophloem phenotype (Extended Data Fig. 4c, d). As previously observed in *brx*, the *pax* PPSE differentiation defects were accompanied by impaired phloem sap delivery into the meristem<sup>16</sup> (Extended Data Fig. 4e). In summary, *pax* mutants represent a (hypomorphic) phenocopy of *brx* mutants.

The *brx* phenotype was not enhanced in *brx pax* double mutants (Extended Data Fig. 4f), suggesting that *brx* is genetically epistatic to *pax*. In turn, the *pax* phenotype was not significantly enhanced by *d6pk/d6pk1* mutations (Extended Data Fig. 4g). However, similar to D6PK or PID, PAX (and PAXL) activated auxin efflux when co-expressed with PINs in *Xenopus laevis* oocytes<sup>12</sup> (Fig. 2d, Extended Data Fig. 4h). However, PAX was the weakest activator in this assay. Moreover, similar to D6PKL proteins, ADP-ribosylation factor–guanine-exchange factor (ARF–GEF) inhibition by brefeldin A (BFA) triggered rapid dissociation of PAX from the plasma membrane (Fig. 3a, Extended Data Fig. 4i). BRX is also BFA-sensitive<sup>18</sup>, yet in direct comparison, BFA-induced BRX plasma-membrane dissociation was slower than for PAX (Fig. 3a, Extended Data Fig. 4j, k). Consistently, BFA treatment also triggered PPSE differentiation defects in a dosage-dependent manner (Extended Data Fig. 4l, m). Moreover, BRX abundance, but not PIN abundance, was severely reduced in *pax* PPSEs (Fig. 3b–g). By contrast, PAX abundance or localization did not substantially depend on BRX (Extended Data Fig. 4n). In protoplasts, BRX localized evenly at the plasma membrane, whereas PAX accumulated in large patches<sup>24</sup> (Extended Data Fig. 4o). Their co-expression recruited BRX into PAX patches. However, a cytoplasmic PAX variant<sup>24</sup> did not disrupt the even plasma-membrane distribution of BRX. These results suggest that PAX is required for efficient BRX plasma-membrane recruitment.

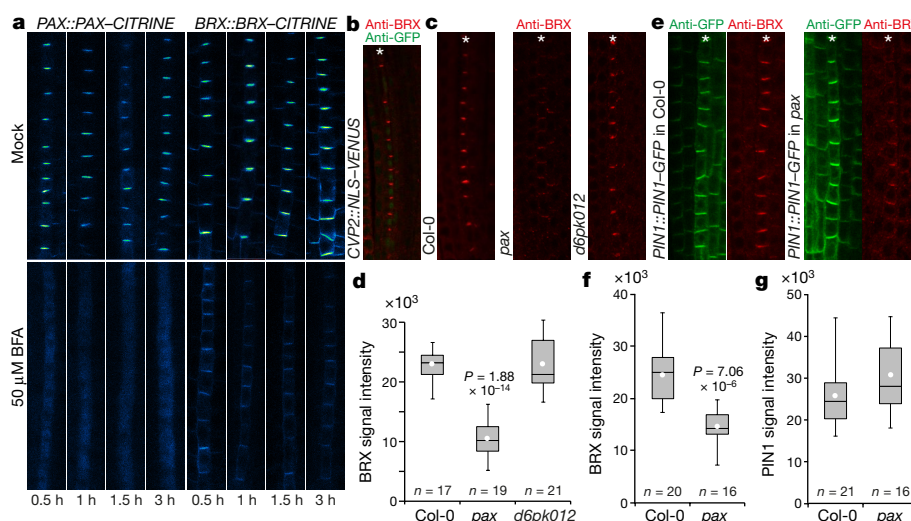
Auxin activity is systemically reduced throughout *brx* root meristems<sup>16</sup>. We thus sought to monitor PIN activity in *brx* or *pax*. We focused our analysis on the dominant PIN in developing PPSEs, PIN1 (Extended

Data Fig. 5a). Corroborating the PIN1–green fluorescent protein (GFP) results, PIN1 abundance and localization in both the protophloem and the meristem were not affected in *pax* or *brx* mutants (Extended



**Fig. 2 | Expression analysis of PAX protein.** **a**, Top, Confocal microscopy of the PAX-CITRINE fusion protein (green fluorescence) expressed under its native promoter in the meristem (longitudinal plane). Bottom, optical cross section. Asterisks indicate PPSE cell files and arrowheads indicate the xylem axis. **b**, Detection of endogenous BRX (red) using anti-BRX antibody staining, or PAX-CITRINE (red) or OPS-CITRINE (green) using anti-GFP antibody staining, in protophloem of fixed meristems (squashed after fixation). Arrows point rootward. **c**, Simultaneous detection of PAX-CITRINE (green) and BRX (red) by immunostaining, demonstrating co-localization. **d**, Auxin transport assays, average retention of radio-labelled auxin in *X. laevis* oocytes expressing the indicated heterologous plant proteins (*n* = 10 per time point; error bars, s.e.m.).





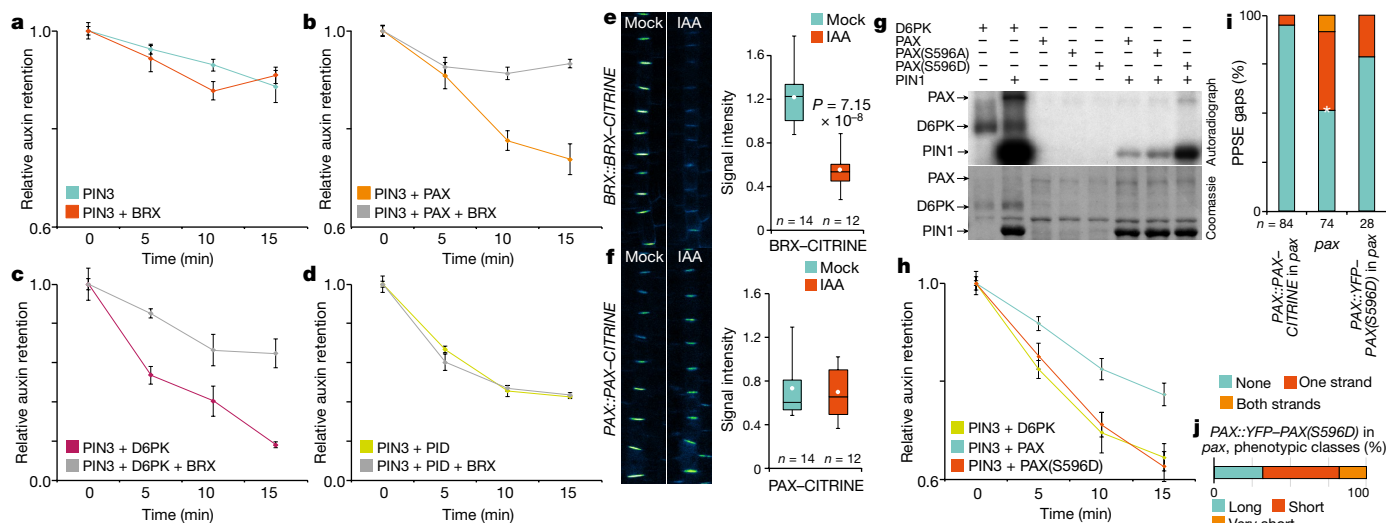
**Fig. 3 | PAX-dependence of efficient BRX plasma-membrane association.** **a**, Plasma-membrane dissociation of PAX-CITRINE and BRX-CITRINE in response to BFA treatment. **b**, Simultaneous detection of nuclear NLS-VENUS protein (green) expressed under the control of PPSE-specific CVP2 promoter and BRX (red) by antibody staining. Asterisks in all microscopy images indicate the PPSE cell file. **c**, Detection of BRX (red) by antibody staining. **d**, BRX signal intensity quantification (anti-BRX antibody detection) in PPSEs (mean from approximately ten

cells per root, arbitrary units). **e**, Simultaneous immunolocalization of PIN1-GFP (anti-GFP, green) expressed under its native promoter and BRX (anti-BRX, red) by antibody staining in Col-0 and *pax*. **f**, **g**, BRX signal intensity quantification (**f**, anti-BRX antibody detection) and PIN1-GFP signal intensity (**g**, anti-GFP antibody detection) in PPSEs (means from approximately ten cells per root, arbitrary units). **d**–**g**, Statistically significant differences from Col-0 are indicated, two-sided Student's *t*-test.

Data Fig. 5b). To survey PIN1 activity, we performed immunostaining with antibodies against PIN1 phosphosites that are critical for PIN1 activation<sup>11,12</sup>. Phosphoserine S231 (J231) signal was significantly reduced in *pax* PPSEs, whereas phosphoserine S271 (J271) was not affected (Extended Data Fig. 5c–e). By contrast, both phosphoserines were barely detectable in *brx* meristems (Extended Data Fig. 5c, f). Reduced PIN1 phosphorylation was also observed in *ops* (Extended Data Fig. 5f), suggesting that meristem-wide reduced PIN1 activity is a secondary systemic consequence of severely disturbed PPSE differentiation, similar to other traits<sup>16</sup>. Yet, *brx* or *pax* protophloem defects

were aggravated in the presence of a *pin1* mutation (Extended Data Fig. 5g–i). In *brx pin1* double mutants, protophloem was frequently barely distinguishable, or even absent (approximately 20% of seedlings) (Extended Data Fig. 5i), underlining the importance of properly regulated auxin transport for PPSE differentiation.

The systemic ramifications of discontinuous protophloem on meristem development can be considered to be a post-catastrophic scenario that is triggered and enhanced by repeated PPSE differentiation failure, and is difficult to recover from once phloem sap (and thus auxin) delivery is impaired<sup>16</sup>. This complicates efforts to untangle cause and



**Fig. 4 | Regulatory input of auxin on PAX and BRX activity.** **a**–**d**, Auxin transport assays in *X. laevis* oocytes expressing the indicated heterologous plant proteins ( $n = 10$  per time point; error bars, s.e.m.). **e**–**f**, Response of BRX (**e**) or PAX (**f**) fusion protein to 5  $\mu$ M auxin (IAA) treatment (3 h), with quantification (arbitrary units; means from approximately ten cells per root). The statistically significant difference is indicated, two-sided Student's *t*-test. **g**, Radioactive in vitro kinase assays with GST fusion proteins of D6PK, PAX or the PAX(S596A) and PAX(S596D) point mutants, with the PIN1 cytosolic loop as substrate (top), and

corresponding loading controls (bottom). **h**, Auxin transport assays in *X. laevis* oocytes expressing the indicated heterologous plant proteins ( $n = 10$  per time point; error bars, s.e.m.). **i**, Quantification of gap-cell frequency in PPSE strands. For PAX::YFP-PAX(S596D), only long root seedlings (see **j**) were scored. *pax* alone was significantly different from PAX::YFP-PAX-CITRINE in *pax* (two-sided Fisher's exact test,  $*P < 0.0001$ ). PAX::YFP-PAX(S596D) was not significantly different from PAX::YFP-PAX-CITRINE in *pax* or from *pax* alone. **j**, Quantification of root phenotypic classes in a PAX::YFP-PAX(S596D) line ( $n = 50$ ).



effect in the cellular action of regulators from the multicellular context. For example, it remained unclear whether *pax* mutants display PPSE differentiation defects because of inefficient BRX plasma membrane recruitment, or whether *brx* mutants display PPSE differentiation defects because of a failure to control PAX activity. To investigate whether BRX interaction with AGC kinases affects auxin transport, we tested the effect of BRX co-expression on kinase-mediated PIN activation in oocytes. In these experiments, BRX substantially inhibited stimulation of auxin efflux by PAX or D6PK (Fig. 4a–c, Extended Data Fig. 5j, k). Because this inhibition was not observed in assays with the more distantly related PID (Fig. 4d), our findings suggest that BRX action affects a subset of related AGC kinases, and that its inhibitory effect is determined by kinase identity.

The observation that BRX inhibits auxin efflux appeared particularly interesting in light of its known auxin-induced plasma-membrane dissociation<sup>18</sup> (Fig. 4e, Extended Data Fig. 6a, b). By contrast, neither PAX abundance nor localization were affected by auxin (Fig. 4f). However, phosphoproteomics indicated auxin-induced phosphorylation of phosphoserine S596 in the PAX activation loop (Extended Data Fig. 6c), which correlated with simultaneously increased PIN1 phosphorylation (Extended Data Fig. 6d). In vitro, recombinant PAX phosphorylated PIN1 with comparably low efficiency, and S596 was dispensable for kinase activation (Fig. 4g, Extended Data Fig. 6e). A PAX(S596D) phosphomimic variant, however, was considerably more active than wild-type PAX and displayed increased phosphorylation activity towards PIN1 (Fig. 4g, Extended Data Fig. 6e). Matching this biochemical observation, PAX(S596D) also stimulated auxin efflux considerably more in oocytes, to a level approximately equal to D6PK (Fig. 4h, Extended Data Fig. 6f). However, unlike wild-type PAX, the PAX(S596D) variant at best partially rescued the *pax* mutant (Fig. 4i, Extended Data Fig. 6g). Moreover, PAX(S596D) frequently triggered a gain-of-function phenotype of even shorter, often barely developing roots (Fig. 4j). Consistent with the D6PK-like activity of PAX(S596D), D6PK expressed from the PAX promoter could not rescue the *pax* phenotype (Extended Data Fig. 6h, i). These findings suggest that fine-tuning of PAX activity is a feature of properly integrated PPSE development.

In a parsimonious interpretation of our results, PAX and BRX act together as a molecular rheostat to modulate auxin efflux dynamically (Extended Data Fig. 7). In this scenario, PAX recruits BRX to the plasma membrane, which inhibits PIN-mediated auxin efflux at lower auxin levels. Because of this inhibition, cellular auxin increases until BRX eventually becomes displaced from the plasma membrane. Concomitantly, PAX is activated and stimulates auxin efflux. Reinforced through auxin-induced BRX transcription<sup>6,18</sup> (Extended Data Fig. 6j, k), this interplay could reach a dynamic steady-state equilibrium, which would impair higher local auxin activity in the multicellular context to properly time PPSE differentiation.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0186-z>.

Received: 14 September 2017; Accepted: 24 April 2018;  
Published online: 06 June 2018

- Benjamins, R. & Scheres, B. Auxin: the looping star in plant development. *Annu. Rev. Plant Biol.* **59**, 443–465 (2008).
- Bliou, I. et al. The PIN auxin efflux facilitator network controls growth and patterning in *Arabidopsis* roots. *Nature* **433**, 39–44 (2005).
- Grieneisen, V. A., Xu, J., Marée, A. F., Hogeweg, P. & Scheres, B. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature* **449**, 1008–1013 (2007).
- Sabatini, S. et al. An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).

- Brunoud, G. et al. A novel sensor to map auxin response and distribution at high spatio-temporal resolution. *Nature* **482**, 103–106 (2012).
- Santuari, L. et al. Positional information by differential endocytosis splits auxin response to drive *Arabidopsis* root meristem growth. *Curr. Biol.* **21**, 1918–1923 (2011).
- Sauer, M. et al. Canalization of auxin flow by Aux/IAA-ARF-dependent feedback regulation of PIN polarity. *Genes Dev.* **20**, 2902–2911 (2006).
- Geldner, N., Friml, J., Stierhof, Y. D., Jürgens, G. & Palme, K. Auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature* **413**, 425–428 (2001).
- Paciorek, T. et al. Auxin inhibits endocytosis and promotes its own efflux from cells. *Nature* **435**, 1251–1256 (2005).
- Barbosa, I. C., Zourelidou, M., Willige, B. C., Weller, B. & Schwechheimer, C. D6 PROTEIN KINASE activates auxin transport-dependent growth and PIN-FORMED phosphorylation at the plasma membrane. *Dev. Cell* **29**, 674–685 (2014).
- Weller, B. et al. Dynamic PIN-FORMED auxin efflux carrier phosphorylation at the plasma membrane controls auxin efflux-dependent growth. *Proc. Natl Acad. Sci. USA* **114**, E887–E896 (2017).
- Zourelidou, M. et al. Auxin efflux by PIN-FORMED proteins is activated by two different protein kinases, D6 PROTEIN KINASE and PINOID. *eLife* **3**, e02860 (2014).
- Rodríguez-Villalón, A. et al. Molecular genetic framework for protophloem formation. *Proc. Natl Acad. Sci. USA* **111**, 11551–11556 (2014).
- Furuta, K. M. et al. *Arabidopsis* NAC45/86 direct sieve element morphogenesis culminating in enucleation. *Science* **345**, 933–937 (2014).
- Kurshumova, W., Smirnova, T., Marcos, D., Zayed, Y. & Berleth, T. Irrepressible *MONOPTEROS/ARF5* promotes *de novo* shoot formation. *New Phytol.* **204**, 556–566 (2014).
- Rodríguez-Villalón, A., Gujas, B., van Wijk, R., Munnik, T. & Hardtke, C. S. Primary root protophloem differentiation requires balanced phosphatidylinositol-4,5-bisphosphate levels and systemically affects root branching. *Development* **142**, 1437–1446 (2015).
- Truernit, E., Bauby, H., Belcram, K., Barthélémy, J. & Palauqui, J. C. OCTOPUS, a polarly localised membrane-associated protein, regulates phloem differentiation entry in *Arabidopsis thaliana*. *Development* **139**, 1306–1315 (2012).
- Scacchi, E. et al. Dynamic, auxin-responsive plasma membrane-to-nucleus movement of *Arabidopsis* BRX. *Development* **136**, 2059–2067 (2009).
- Breda, A. S., Hazak, O. & Hardtke, C. S. Phosphosite charge rather than shootward localization determines OCTOPUS activity in root protophloem. *Proc. Natl Acad. Sci. USA* **114**, E5721–E5730 (2017).
- Mouchel, C. F., Osmont, K. S. & Hardtke, C. S. BRX mediates feedback between brassinosteroid levels and auxin signalling in root growth. *Nature* **443**, 458–461 (2006).
- Kondo, Y. et al. Vascular cell induction culture system using *Arabidopsis* leaves (VISUAL) reveals the sequential differentiation of sieve element-like cells. *Plant Cell* **28**, 1250–1262 (2016).
- Galván-Ampudia, C. S. & Offringa, R. Plant evolution: AGC kinases tell the auxin tale. *Trends Plant Sci.* **12**, 541–547 (2007).
- Willige, B. C. et al. D6PK AGCVIII kinases are required for auxin transport and phototropic hypocotyl bending in *Arabidopsis*. *Plant Cell* **25**, 1674–1688 (2013).
- Barbosa, I. C. et al. Phospholipid composition and a polybasic motif determine D6 PROTEIN KINASE polar association with the plasma membrane and tropic responses. *Development* **143**, 4687–4700 (2016).

**Acknowledgements** We thank the University of Lausanne Protein Analysis Facility for mass spectrometry services, the Swiss National Science Foundation for Grant 31003A\_166394 (C.S.H.), the German-Israeli Foundation for I-236-203.17-2014 (C.S.) and the Deutsche Forschungsgemeinschaft for SCHW751/12-2 (C.S.), HA 3468/6-1 and SFB924 (U.Z.H.).

**Reviewer information** Nature thanks A. P. Mahonen, D. Weijers and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** P.M., A.E.L.B., M.Z., M.K., B.M., A.F., W.X.S. and P.C. performed experiments and analysed data. P.M., A.E.L.B., M.Z., W.X.S., U.Z.H., C.S. and C.S.H. designed experiments. P.M., A.E.L.B., U.Z.H., C.S. and C.S.H. wrote the manuscript.

**Competing interests** The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0186-z>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0186-z>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to C.S. and C.S.H.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment. Experiments were repeated two to four times. All attempts at replication were successful.

**Plant materials and growth conditions.** The wild-type *Arabidopsis* line used in this study was the *A. thaliana* L. Heynh reference accession Columbia-0 (Col-0), which was also the genetic background for the mutants and transgenic lines. For plant tissue culture, seeds were surface-sterilized, stratified for 2 days in the dark at 4 °C, and germinated in vertically placed Petri dishes on 0.9% agar and 0.5 × Murashige and Skoog (1/2 MS) medium (Duchefa) with 0.3% sucrose at 22 °C under continuous light. The mutant *pax-1* and *pax-2* alleles (T-DNA insertion lines SAIL\_688\_B04 and GABI\_274F04, respectively) were identified from available collections and obtained from the Nottingham *Arabidopsis* Stock Centre. The following transgenic and mutant lines have been described elsewhere: *BRX::BRX-CITRINE*<sup>13</sup>, *PIN1::PIN1-GFP*<sup>25</sup>, *PIN3::PIN3-GFP*<sup>26</sup>, *PIN7::PIN7-GFP*<sup>27</sup>, *CVP2::NLS-VENUS*<sup>16</sup>, *35S::DII-NLS-VENUS* and *35S::mDII-NLS-VENUS*<sup>5</sup>, *brx-2*<sup>13</sup>, *ops-2*<sup>17</sup>, *pin1-613*<sup>28</sup>, *aux1-729*, *d6pk*, *d6pk11*, *d6pk12*, *d6pk13*, as well as their *d6pk012* triple and *d6pk0123* quadruple mutants<sup>12</sup>. Primers used for genotyping are summarized in Extended Data Table 1.

**Constructs and generation of transgenic lines.** Transgenes for plant transformation were created in suitable binary vectors and produced using standard molecular biology procedures and/or the NEBuilder HiFi DNA Assembly Reaction Protocol. The *35S::mDII-VENUS* and *35S::DII-VENUS* lines in Col-0 and *brx-2* backgrounds have previously been described<sup>6</sup>. For the *CVP2::MP<sup>A</sup>* or *CLE45::MP<sup>A</sup>* constructs, a 2.6-kb genomic promoter fragment upstream of the initiation codon of the *CVP2* gene<sup>16</sup> or a 2.0-kb genomic promoter fragment upstream of the initiation codon of the *CLAVATA3/EMBRYO SURROUNDING REGION 45 (CLE45)* gene<sup>16</sup> was amplified, combined with amino acids 1–794 of the *MP* open reading frame<sup>15</sup> and introduced into the pCambia1305.1 binary vector. To generate the translational *PAX::PAX-CITRINE* fusion, the *PAX* promoter region (4.5-kb upstream of the ATG start codon) was amplified and cloned into pDONR P4P1R as well as a genomic fragment of the *PAX* transcript region without a STOP codon into pDONR 221. The entry clones together with CITRINE in pDONR P2RP3 were cloned into the destination vector pH7m34GW by the multisite Gateway recombination system. To create *UBQ10::PAX-CITRINE*, the entry clones containing the *UBQ10* promoter in pDONR P4P1R, the *PAX* coding sequence without a STOP codon in pDONR 221, and the CITRINE coding sequence in pDONR P2RP3 were combined into binary vector pH7m34GW. The binary constructs were introduced into *Agrobacterium tumefaciens* strain GV3101<sup>pMP90</sup> and transformed into the pertinent *Arabidopsis* genotypes using the floral dip method. For recombinant expression of the glutathione-S-transferase N-terminally tagged fusion proteins GST-PAX, GST-PAX(S596D) and GST-PAX(S596A), a Gateway-compatible attB-flanked PCR product of *PAX* coding sequence was amplified from cDNA and cloned into the Gateway-compatible donor vector pDONR 201 (Invitrogen). The mutagenesis leading to *PAX*(S596D) phosphomimetic or *PAX*(S596A) phospho-mutant variants in the activation loop of the *PAX* coding sequence was achieved using site-directed mutation PCR on a pDONR 201 entry clone carrying the *PAX* coding sequence insert. The resulting pDONR 201 entry clones served as substrate to recombine the *PAX* coding sequences into the pDEST15 (Life Technologies) destination vector that was ultimately used for recombinant protein expression. The expression vectors pDEST15 containing GST-D6PK and GST-PIN1 cytosolic loop have previously been described<sup>12</sup>. Primers used for cloning are summarized in Extended Data Table 1.

**Microscopy.** To visualize reporter genes and staining signals, fluorescence for CITRINE (excitation 514 nm, emission 529 nm), VENUS (excitation 515 nm, emission 528 nm), propidium iodide (PI) (excitation 536 nm, emission 617 nm), Alexa Fluor 488 (excitation 498 nm, emission 520 nm) and Alexa Fluor 546 (excitation 556 nm, emission 573 nm) were detected in seedlings examined under Zeiss LSM 700 or 710 inverted confocal scanning microscopes. Pictures were taken with 20× or 40× water/oil immersion objectives. PI staining of seven-day-old seedlings was used for quantification of protophloem cell size. For presentation, composite images had to be assembled in various instances. Sequential scanning was used for co-localization studies to avoid any interference between fluorescence channels. For image analyses, ImageJ (NIH; <https://rsb.info.nih.gov/ij/>), Zeiss Zen (black edition) and Imaris software were used. If necessary, images were processed using the 'sharpen' tool for clearer visualization of cellular organization. For signal quantifications, all samples were analysed in the same area of the root meristem, and the average signal intensity per transgenic line was calculated as the mean of means. Statistical significance was evaluated with Student's *t* test.

**Pharmacological and hormonal treatments.** For treatments, 5–7-day-old seedlings were either grown on, or transferred either onto solid or into liquid 1/2 MS

medium with or without the chemicals and incubated for the indicated time. Drugs and hormones used were as follows: BFA (dissolved in DMSO), IAA (dissolved in DMSO), L-kynurenine (dissolved in DMSO), PI (1 mg ml<sup>-1</sup> in water, diluted 1:25). **VISUAL assay and proteomics.** The VISUAL protocol<sup>21</sup> was performed as previously described<sup>19</sup> with subsequent BRX-CITRINE or YFP pull down. In brief, cotyledons of six-day-old transgenic *Arabidopsis* seedlings were cultured for 3 days in induction medium and subsequently ground in extraction buffer. Supernatants (4 mg of total protein extract in two technical replicates) were incubated with anti-GFP beads (GFP-Trap\_MA, Chromotek). Beads were magnetically separated and protein was eluted in 2× SDS sample buffer, loaded on an SDS-PAGE gel for electrophoresis, and subsequently analysed by liquid chromatography with tandem mass spectrometry.

**Protein immunolocalization.** Whole mount immunolocalization on five-day-old seedlings was performed as described previously<sup>7</sup>. The primary antibody dilutions were: 1:600 for anti-GFP mouse (Roche), 1:500 for anti-BRX rabbit (this study), 1:100 for anti-PIN1 S1-P rabbit<sup>11</sup>, 1:300 for anti-PIN1 S4-P rabbit<sup>11</sup> and 1:100 for anti-PIN1 goat (Santa Cruz Biotechnology). The secondary antibody dilutions were: 1:600 for Alexa Fluor 488 anti-mouse (Molecular Probes) and 1:600 for Alexa Fluor 546 anti-rabbit (Molecular Probes). The anti-BRX antibody was obtained by custom antibody production directed against the keyhole-limpet hemocyanin (KLH)-conjugated BRX peptide GGSSNYGPGSYHGCG with affinity purification (Agrisera).

**Oocyte experiments.** Auxin transport assays in *X. laevis* oocytes were carried out as described<sup>12,30</sup>. The oocytes were obtained from the animal facility of the Technical University of Munich, Department of Nutritional Physiology. The animals were kept in accordance with local guidelines and regulations. To monitor expression levels, post-assay immunoblots were performed with anti-PIN1 sheep and anti-PIN3 sheep primary antibodies (Nottingham *Arabidopsis* Stock Centre, used at 1:5,000 dilution), and anti-GFP rabbit (custom antibody<sup>31</sup>, 1:2,000 dilution). The secondary antibodies were anti-sheep from donkey (1:5,000, Santa Cruz Biotechnology) and anti-rabbit from goat (1:10,000, Santa Cruz Biotechnology).

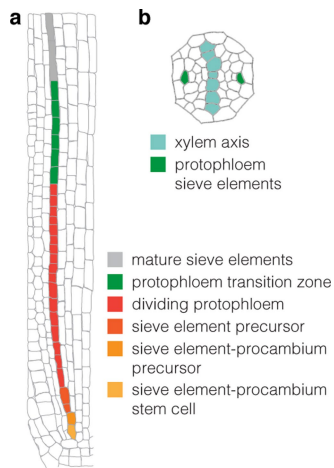
**Protoplast transformation.** Protoplasts were isolated from *Arabidopsis* suspension-cultured cells (Col-0) five to seven days after subculturing by incubation of 2 g of cell culture with 1% Cellulase R-10 (SERVA) and 0.25% Macerozyme R-10 (SERVA). Typically, protoplasts were transformed with 20 µg of plasmid DNA using polyethylene glycol-mediated transformation and analysed after 16 to 20 h of incubation as described<sup>24</sup>.

**In vitro kinase assay.** The in vitro kinase assay was performed using recombinant glutathione-S-transferase N-terminally tagged fusion proteins GST-PIN1<sup>12</sup> CL (cytosolic loop), GST-D6PK<sup>12</sup>, GST-BRX<sup>18</sup>, GST-PAX, GST-PAX(S596D) and GST-PAX(S596A), expressed in the *Escherichia coli* strain BL21(DE3) and purified using Glutathione Sepharose 4B (GE Healthcare). The kinase reactions were performed by incubating the purified GST-fusion proteins for 60 min at 28 °C in the kinase reaction buffer (25 mM Tris HCl pH 7.5, 5 mM MgCl<sub>2</sub>, 0.2 mM EDTA, 1× cComplete protease inhibitor cocktail (Roche)), supplemented with 10 µCi [γ-<sup>32</sup>P] ATP (370 Mbq, specific activity 185 Tbq, Hartmann Analytics). The reactions were stopped by boiling the protein samples mixed with 5× concentrated Laemmli buffer for 10 min. Subsequently, the protein mixtures were separated by SDS-PAGE. After samples had been run, the SDS-PAGE gel was vacuum-dried and used for autoradiography. The same gel was later rehydrated and stained with Coomassie Brilliant Blue to serve as a loading control.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

**Data availability.** The datasets displayed in the current study are available from the corresponding authors upon reasonable request. For gel source images, see Supplementary Fig. 1.

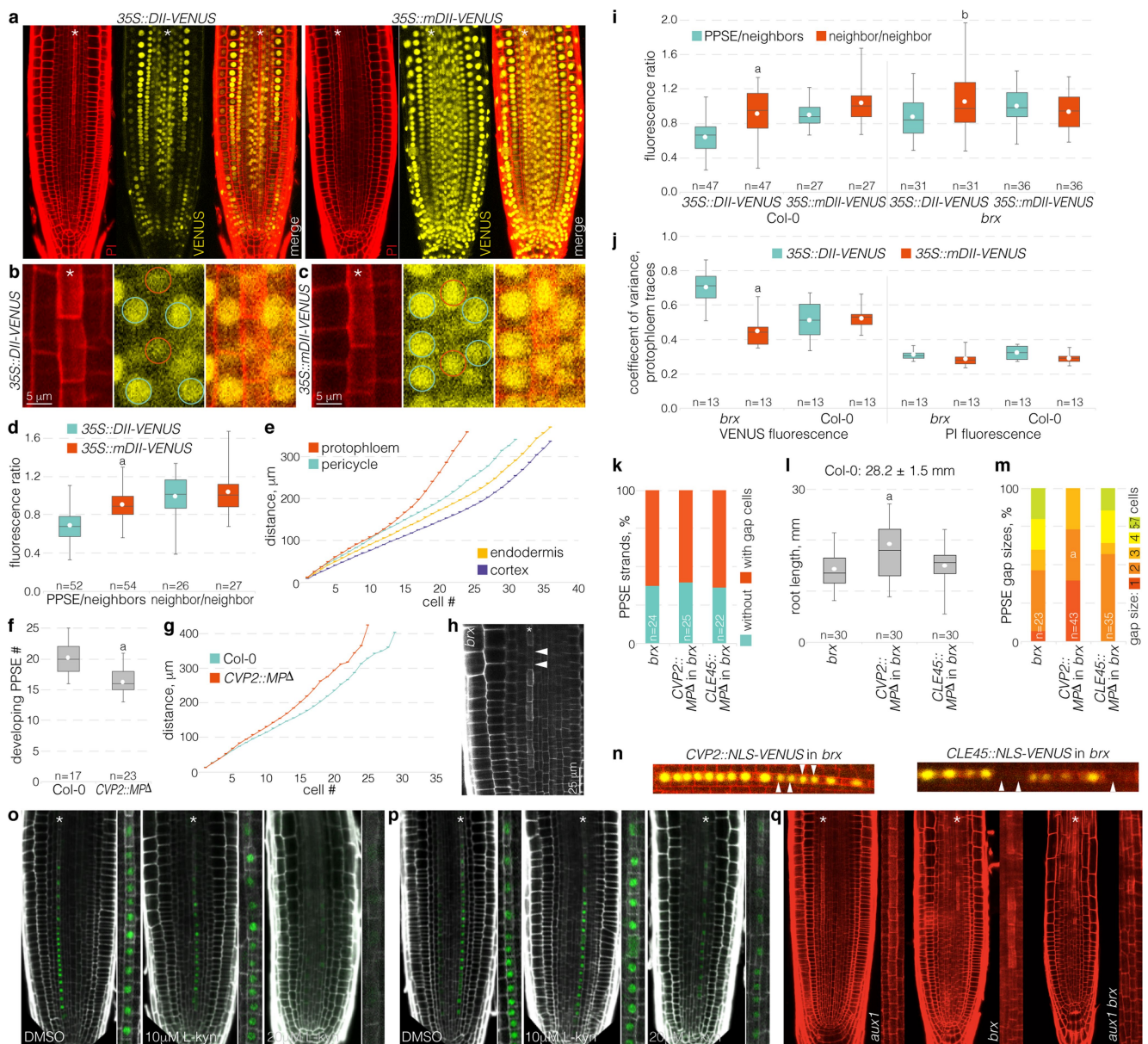
25. Benková, E. et al. Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115**, 591–602 (2003).
26. Zádorníková, P. et al. Role of PIN-mediated auxin efflux in apical hook development of *Arabidopsis thaliana*. *Development* **137**, 607–617 (2010).
27. Friml, J. et al. Efflux-dependent auxin gradients establish the apical-basal axis of *Arabidopsis*. *Nature* **426**, 147–153 (2003).
28. Bennett, T. et al. The *Arabidopsis* MAX pathway controls shoot branching by regulating auxin transport. *Curr. Biol.* **16**, 553–563 (2006).
29. Marchant, A. et al. AUX1 promotes lateral root formation by facilitating indole-3-acetic acid distribution between sink and source tissues in the *Arabidopsis* seedling. *Plant Cell* **14**, 589–597 (2002).
30. Fastner, A., Absmanner, B. & Hammes, U. Z. Use of *Xenopus laevis* oocytes to study auxin transport. *Methods Mol. Biol.* **1497**, 259–270 (2017).
31. Absmanner, B., Stadler, R. & Hammes, U. Z. Phloem development in nematode-induced feeding sites: the implications of auxin and cytokinin. *Front. Plant Sci.* **4**, 241 (2013).



**Extended Data Fig. 1 | Overview of protophloem development.**

**a**, Illustration of protophloem development from the stem cell to the mature sieve element in the *Arabidopsis* root meristem. **b**, Illustration of a cross section through the stele of an *Arabidopsis* root meristem, highlighting the arrangement of the two sieve element strands and the xylem axis.



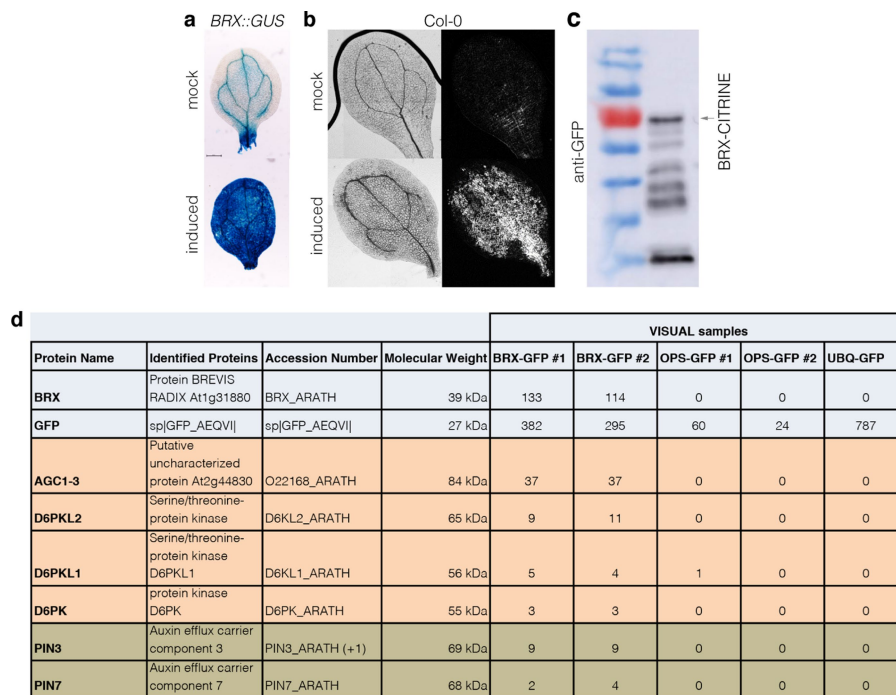


Extended Data Fig. 2 | See next page for caption.



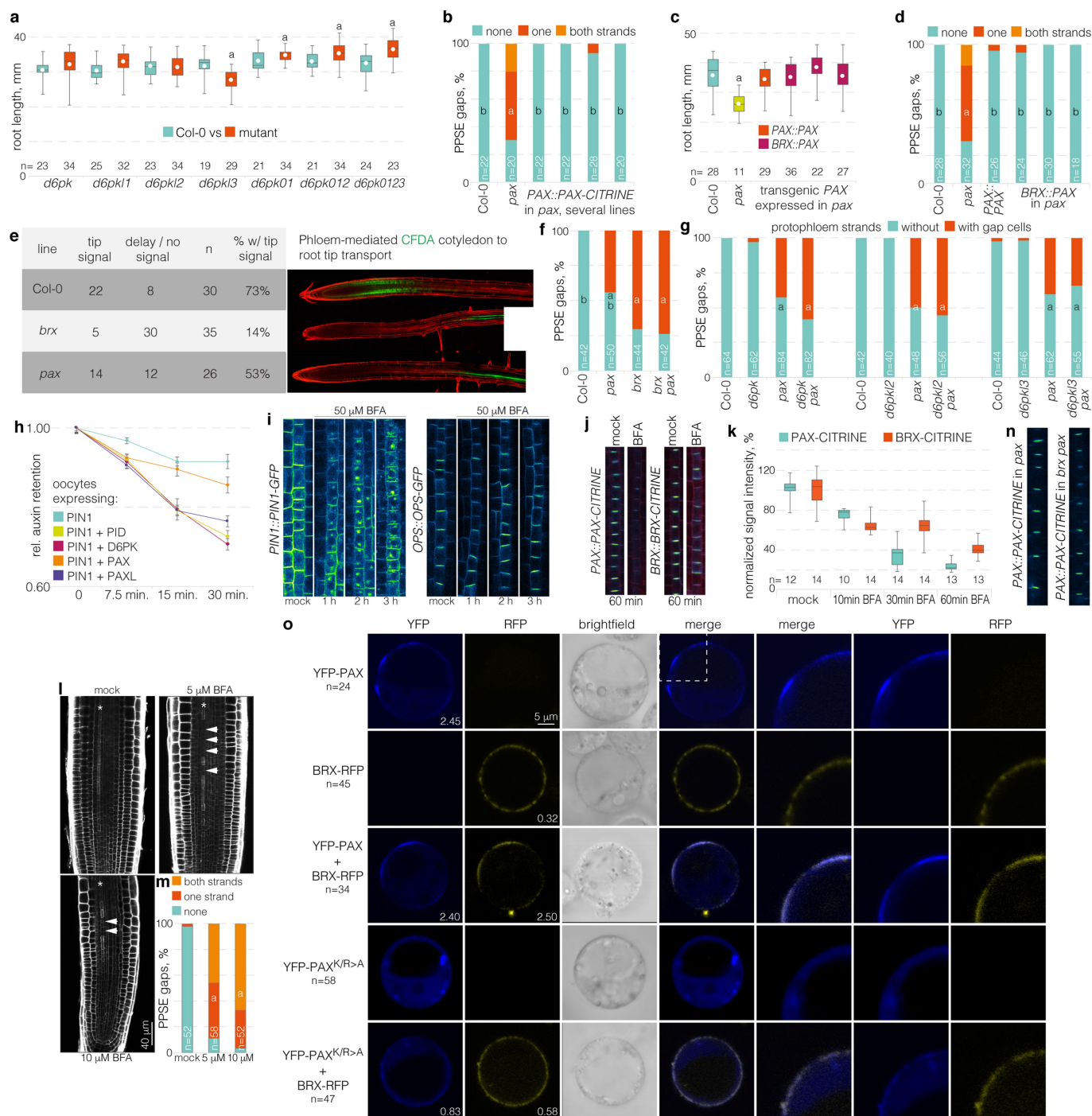
**Extended Data Fig. 2 | Auxin activity in developing PPSEs.** **a**, Confocal microscopy of the inverse auxin activity reporter DII-VENUS and its negative control mDII-VENUS (yellow fluorescence) in the root meristem (PI staining, red) of wild-type Col-0 plants. Asterisks indicate sieve element cell files. **b**, Confocal microscopy of constitutively expressed DII-VENUS in developing PPSEs and neighbouring cell files. Left, PI cell-wall staining (red); middle, DII-VENUS fluorescence (yellow; PPSE nuclei marked with red circles, nuclei in neighbouring cell files with blue circles); right, overlay. **c**, As in **b**, for mDII-VENUS. **d**, Relative intensity of the DII-VENUS reporter and its mDII-VENUS control in the nuclei of Col-0 PPSEs as compared to the nuclei of directly neighbouring cells. The statistically significant difference between DII-VENUS and mDII-VENUS in the PPSE/neighbours group is indicated (two-sided Student's *t*-test; *a*,  $P = 5.86 \times 10^{-11}$ ). **e**, Cumulative average cell length in different root cell files, starting from the respective first stem-cell daughters (cell #1) ( $n = 11$  wild-type Col-0 roots). **f**, Number of developing PPSEs from the first stem-cell daughter up to the first transition zone PPSE (protophloem length) in seven-day-old Col-0 seedlings, and transgenic seedlings expressing a constitutively active derivative of the auxin response factor MONOPTEROS ( $MP^{\Delta}$ ) under control of the PPSE-specific *CVP2* promoter. *a*,  $P = 3.16 \times 10^{-6}$ ; two-sided Student's *t*-test. **g**, Cumulative average cell length in the developing protophloem, starting from the first stem-cell daughter (cell #1) ( $n = 23$  each). Elongation occurs prematurely in *CVP2::MP^{\Delta}* plants. **h**, Confocal microscopy of a *brx* root meristem, focused on one of the sieve element strands (asterisk). Arrowheads point out gap cells, which fail to build up the characteristic PPSE cell wall owing to a failure to differentiate. **i**, Relative intensity of the DII-VENUS reporter and its mDII-VENUS control in the nuclei of Col-0 and *brx* PPSEs as

compared to nuclei of cells in directly neighbouring files. Statistically significant differences between PPSE/neighbours and neighbour/neighbour in the Col-0 and *brx* DII-VENUS groups are indicated (two-sided Student's *t*-test; *a*,  $P = 2.49 \times 10^{-7}$ ; *b*,  $P = 0.026$ ). **j**, Coefficient of variance for fluorescence traces of the DII-VENUS reporter and its mDII-VENUS control (left) and PI staining (right) along protophloem cell files. The statistically significant difference in VENUS fluorescence in the *brx* group is indicated (two-sided Student's *t*-test; *a*,  $P = 2.30 \times 10^{-7}$ ). **k**, Quantification of PPSE strands with gaps in roots of indicated genotypes. **l**, Root length in seven-day-old seedlings for indicated genotypes. The statistically significant differences between *CVP2::MP^{\Delta}* in *brx* and *brx* alone ( $P = 0.0017$ ) and between *CVP2::MP^{\Delta}* in *brx* and *CLE45::MP^{\Delta}* in *brx* ( $P = 0.0052$ ) are indicated by the character *a*. **m**, Distribution of gap size in protophloem strands of seven-day-old seedlings with gaps of indicated genotypes. The statistically significant differences between *CVP2::MP^{\Delta}* in *brx* and *brx* alone ( $P = 0.0008$ ) and between *CVP2::MP^{\Delta}* in *brx* and *CLE45::MP^{\Delta}* in *brx* ( $P = 0.0051$ ) are indicated by the character *a* (two-sided  $\chi^2$  test). **n**, Expression of fluorescent NLS-VENUS reporter in PPSEs of *brx* mutants, driven by either *CVP2* or *CLE45* promoter. Arrowheads indicate gap cells. **o**, **p**, Expression of *CVP2::NLS-VENUS* reporter (green fluorescence) in PPSE cell files (asterisks) of six-day-old Col-0 root meristems (PI staining, white) grown in the presence of (**o**), or transferred for 48 h onto (**p**), increasing amounts of the auxin biosynthesis inhibitor L-kynurenine (L-kyn). On the higher concentration, PPSE cell files (magnified) were barely distinguishable. **q**, Confocal microscopy of seven-day-old root meristems (PI staining, red). Asterisks indicate sieve element cell files (magnified, barely distinguishable in *aux1 brx*).



**Extended Data Fig. 3 | Identification of BRX interactors.** **a**, Induction of *BRX* expression in cotyledons in the VISUAL transdifferentiation assay, as indicated by a *BRX::GUS* reporter gene. **b**, Visualization of successful tracheary element differentiation using polarized light

microscopy. **c**, Western analysis of BRX–CITRINE fusion protein after immunoprecipitation. **d**, List of the top BRX interactors, indicating the number of peptides isolated as compared to controls.

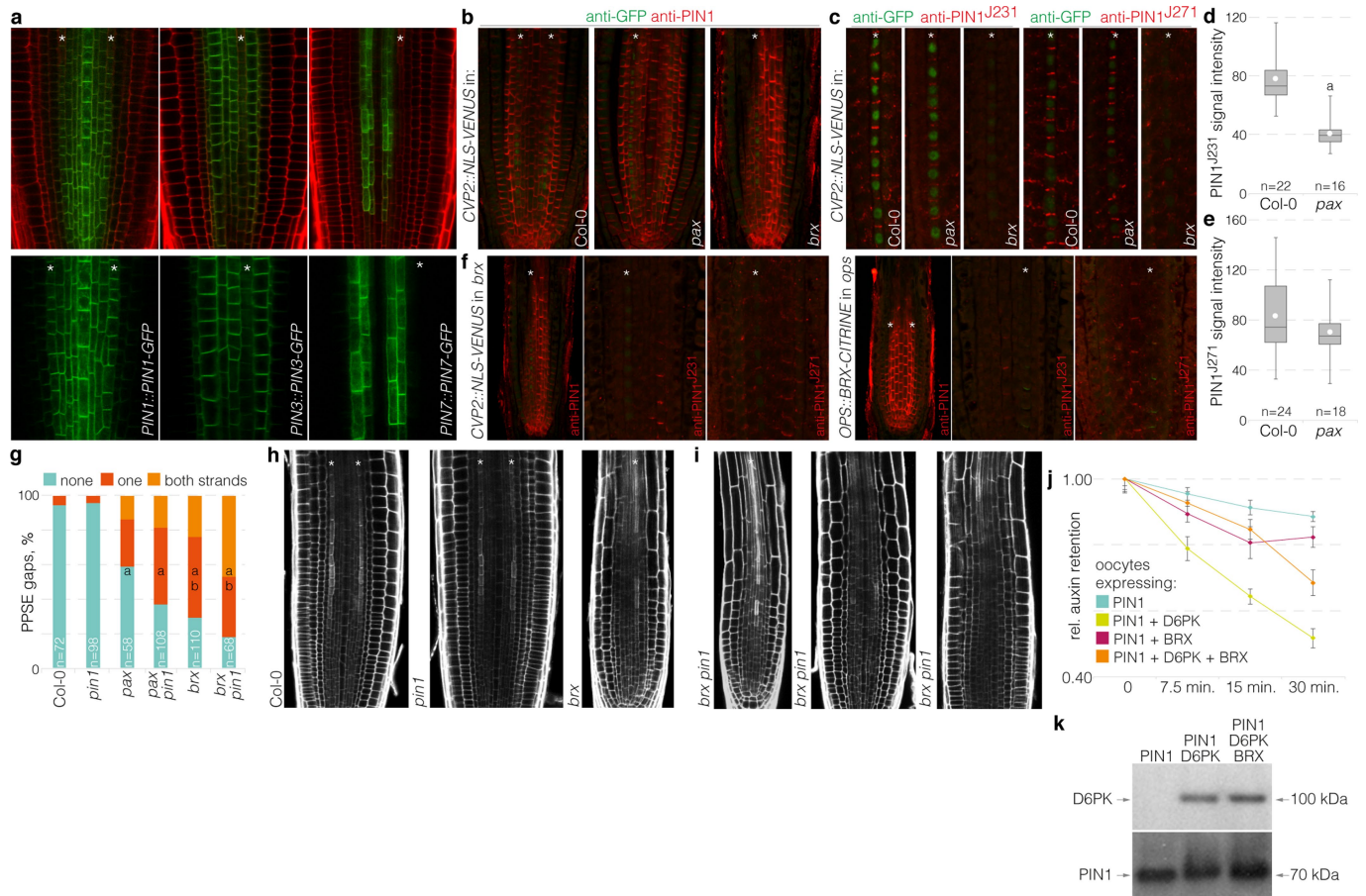


Extended Data Fig. 4 | See next page for caption

**Extended Data Fig. 4 | Phenotypic analysis of *pax*-related mutants and transgenic lines.** **a**, Root length in seven-day-old seedlings for indicated mutants and parallel Col-0 controls. Statistically significant differences between Col-0 and mutants are indicated (Student's *t*-test, two-sided; *a*,  $P < 0.02$ ). **b**, Quantification of gap-cell frequency in protophloem strands of six-day-old seedlings. Statistically significant differences are indicated (two-sided Fisher's exact test; *a*, *pax* versus Col-0; *b*, others versus *pax*; all  $P$  values  $< 0.001$ ). **c**, Root length in seven-day-old seedlings for Col-0, *pax* and transgenic lines in the *pax* mutant background that expressed PAX under the control of its native promoter or the *BRX* promoter. The statistically significant difference between *pax* and Col-0 is indicated (two-sided Student's *t*-test; *a*,  $P = 0.00016$ ). **d**, Quantification of gap-cell frequency in protophloem strands of six-day-old seedlings. Statistically significant differences are indicated (two-sided Fisher's exact test; *a*, *pax* versus Col-0; *b*, others versus *pax*; all  $P$  values  $< 0.001$ ). **e**, Phloem-mediated translocation of carboxyfluorescein diacetate succinimidyl ester (CFDA) dye (green fluorescence) into the phloem-unloading zone of the root tip 45 min after CFDA application to the cotyledons of four-day-old seedlings, and corresponding classification of CFDA signal at the end of the experiment. **f**, Quantification of gap-cell frequency in protophloem strands of six-day-old seedlings. Statistically significant differences are indicated (two-sided Fisher's exact test; *a*, others versus Col-0; *b*, Col-0 and *pax* versus *brx*; all  $P$  values  $< 0.01$ ). **g**, Quantification of gap-cell frequency in protophloem strands of six-day-old seedlings. Statistically significant

differences are indicated (two-sided Fisher's exact test; *a*, others versus Col-0, all  $P$  values  $< 0.001$ ). **h**, Auxin transport assays performed in *X. laevis* oocytes expressing the indicated heterologous plant proteins ( $n = 10$  oocytes per time point; error bars, s.e.m.). **i**, BFA control experiments. Accumulation of PIN1-GFP fusion protein in BFA compartments (left), and comparative BFA insensitivity of OPS-GFP fusion protein (right). **j**, Dissociation of PAX-CITRINE and BRX-CITRINE fusion proteins from the plasma membrane in response to 5  $\mu$ M BFA treatment. **k**, Quantification of PAX-CITRINE and BRX-CITRINE fluorescence signal at the plasma membrane in response to 5  $\mu$ M BFA treatment, normalized to allow direct comparison (means of approximately ten cells per root). **l**, Confocal microscopy of six-day-old PI-stained root meristems grown on mock or low BFA concentration as indicated. Asterisks indicate PPSE cell files and arrowheads indicate gap cells. **m**, Quantification of gap-cell frequency in PPSE strands of roots shown in (**l**). Statistically significant differences are indicated (two-sided Fisher's exact test; *a*, others versus mock,  $P < 0.0001$ ). **n**, Expression of PAX-CITRINE fusion protein under its native promoter, in *pax* single or *brx pax* double mutants. **o**, Transient expression of the indicated fusion proteins, alone or in combination, in *Arabidopsis* protoplasts. The PAX<sup>K/R>A</sup> variant carries point mutations in a polybasic stretch that is required for plasma membrane interaction<sup>24</sup>. The average number of patches per protoplast is indicated.

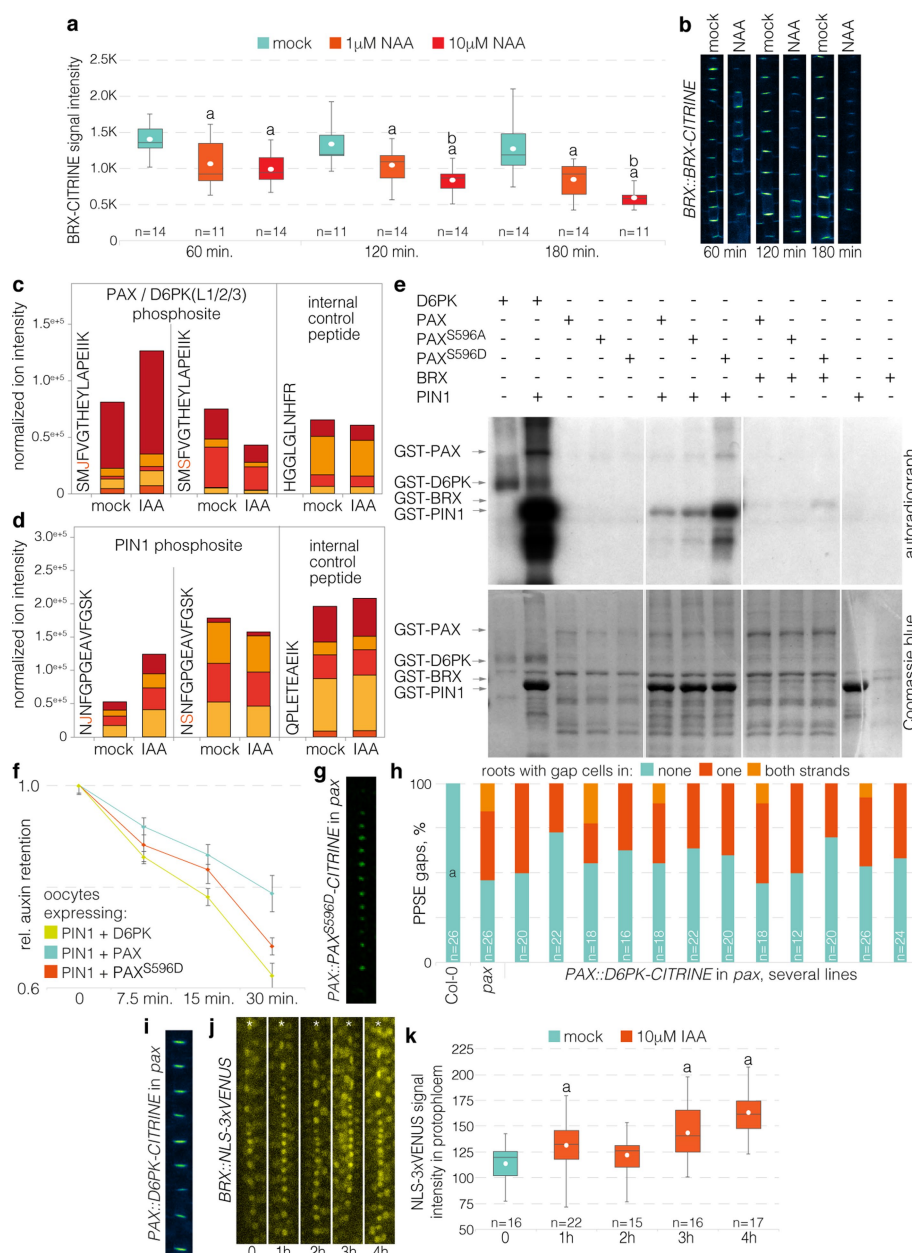




### Extended Data Fig. 5 | PIN activity in the root protophloem.

**a**, Confocal microscopy of indicated reporter genes (green fluorescence) in the root meristem (PI staining, red) of Col-0 wild-type plants (top), and magnification without PI background (bottom). Asterisks indicate sieve element cell files. **b**, Immunolocalization of nuclear localized NLS-VENUS (green) expressed under control of PPSE-specific CVP2 promoter, and PIN1 (red) by antibody staining. Asterisks indicate PPSE cell files. **c**, Simultaneous immunolocalization of CVP2-driven NLS-VENUS (green) with different anti-PIN1 antibodies that specifically detect phosphorylated PIN1 residues S231 (J231) or S271 (J271). **d**, Quantification of the J231 phosphosite signal intensity (means from approximately ten cells per root, arbitrary units). The statistically significant difference is indicated (two-sided Student's *t*-test; a,  $P = 1.2 \times 10^{-9}$ ). **e**, Quantification of the J271 phosphosite signal intensity (means from approximately ten cells per root, arbitrary units). **f**, Immunolocalization of PIN1, and the J231 and J271 PIN1 phosphosites (red) in *brx* (left) or *ops* (right) by antibody staining, with an OPS::BRX-CITRINE

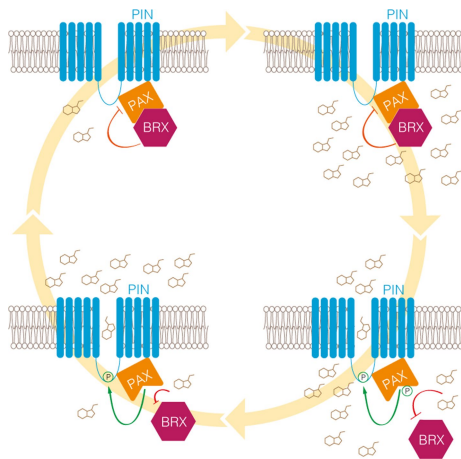
or CVP2::NLS-VENUS reporter in the background for the identification of PPSE cell files (asterisks). **g**, Quantification of gap-cell frequency in protophloem strands of six-day-old seedlings for the indicated genotypes. Statistically significant differences are indicated (two-sided Fisher's exact test; a, Col-0 and *pin1* versus others,  $P < 0.0001$ ; b, *brx* or *pax* single mutant versus *brx pin1* or *pax pin1* double mutants,  $P < 0.02$ ). **h**, Confocal microscopy of representative six-day-old Col-0, *pin1*, and *brx* root meristems (PI staining, white). Asterisks indicate PPSE cell files. **i**, Different phenotypic classes occurring in *brx pin1* double mutant root meristems (PI staining, white). PPSE cell files were frequently barely distinguishable or missing. **j**, Auxin transport assays performed in *X. laevis* oocytes expressing the indicated heterologous plant proteins ( $n = 10$  oocytes per time point; error bars, s.e.m.). **k**, Western blot analysis of the oocytes used in j, demonstrating that BRX expression does not interfere with D6PK or PIN1 expression or stability (detection of YFP-D6PK and PIN1 with anti-GFP and anti-PIN1 antibodies, respectively).



### Extended Data Fig. 6 | BRX auxin response and PAX specificity.

**a, b**, Response of BRX–CITRINE fusion protein to treatment with 1  $\mu$ M or 10  $\mu$ M auxin (NAA), time course experiment (**b**) with quantification (**a**, means from approximately ten cells per root, arbitrary units). Statistically significant differences are indicated (two-sided Student's *t*-test; **a**, mock versus others,  $P < 0.0094$ ; **b**, 1  $\mu$ M versus 10  $\mu$ M auxin,  $P < 0.0028$ ). **c**, Phosphoproteomics of auxin-treated seedlings, showing normalized abundance of a conserved phosphosite in PAX, D6PK, D6PKL1-3, and AGC1-6, with subfragments indicated in different colours. **d**, Same as **c**, for a PIN1 phosphosite. **e**, Radioactive in vitro kinase assays with GST fusion proteins of D6PK, PAX, or PAX(S596A) and PAX(S596D) point mutants, with BRX or the PIN1 cytosolic loop as substrate (top) and corresponding loading controls (bottom). **f**, Auxin transport assays performed in *X. laevis* oocytes expressing the indicated heterologous

plant proteins ( $n = 10$  oocytes per time point; error bars, s.e.m.). **g**, Polar localization of the YFP–PAX(S596D) variant in developing PPSEs of a *pax* mutant. **h**, Quantification of gap-cell frequency in protophloem strands of seven-day-old *pax* mutant seedlings that express a D6PK–CITRINE fusion protein under the control of the *PAX* promoter. The statistically significant difference is indicated (two-sided Fisher's exact test; **a**, Col-0 versus *pax* and transgenic lines,  $P < 0.0001$ ). **i**, Polar localization of D6PK–CITRINE fusion protein in developing PPSEs of a *pax* mutant. **j, k**, Auxin induction of *BRX* transcription in developing PPSE cell files (asterisks) visualized using an NLS–3 $\times$ VENUS reporter gene (**j**), with corresponding quantification of nuclear fluorescence signal (**k**). Statistically significant differences are indicated (one-sided Student's *t*-test; **a**, versus preceding time point,  $P < 0.0153$ ).



**Extended Data Fig. 7 | Molecular rheostat model for PAX–BRX action in the regulation of auxin efflux.** Proposed model for the cellular action of PAX and BRX as elements of a molecular rheostat. BRX interacts with PAX at the plasma membrane, where it inhibits PIN-mediated auxin efflux at lower auxin levels. Because of reduced PIN-mediated auxin efflux, cellular auxin levels increase so that, eventually, BRX becomes displaced from the plasma membrane. Concomitantly, PAX becomes activated and increasingly stimulates auxin efflux. Reinforced through auxin-induced *BRX* transcription and decreasing cellular auxin levels, BRX can return to the plasma membrane and again inhibit auxin efflux. This interplay would lead to a dynamic steady-state equilibrium that fine-tunes auxin levels along a cell file.

**Extended Data Table 1 | List of oligonucleotides used in this study****Genotyping T-DNA insertion lines 5'→3'**

LB3 Sail	TAGCATCTGAATTTTCATAACCAATCTCGATACAC
LB Gabi o8474	ATAATAACGCTGCGGACATCTACATTTT
LBb1.3 Salk	ATT TTG CCG ATT TCG GAA C
brx-2_F417	GTCAGTGTTTGCTTCCTCTCTATG
brx-2_R650	TATTTCCCTTGTCTAGGTAAGAATCC
brx-2_ski3	TGATCCATGTAGATTTCCCGGACATGAA
pax_Sail_RP	ATAGTAGCGGCCTAAGCGAAG
pax_Sail_LP	CTGCATAGCTAGTTGCTGGTG
pax_Sail_LP_UTR	CTACTACGGAACCAAATCTTGG
pax_GK_LP	GTACTCCGAGAAATGCTTCCC
pax_GK_RP	AAATTGGAGCTGCTGATGATG
d6pk LP	CAA GAT CAA CCG GTT TAG GAA TCT
d6pk RP	GGC TCA AAC TGA AAG AGA GAT ATT GC
d6pkI1 LP	CATTTCCATGGAAGGAAGGTGATGAGCT
d6pkI1 RP	ACTCCAGAACCATACTTCGAGGCCAT
d6pkI2 LP	CTTCGCCTTTGATGATCTCTG
d6pkI2 RP	AGTGACGAGAGTAGCTGCAGC
d6pkI3 LP	CGGACACTAATCGGATCGGA
d6pkI3 RP	GGCAAGAAAGGATGATCTAACG
pin1_SALK047613_F	CAAAAACACCCCCAAAATTTCTCT
pin1_SALK047613_RP	AATCATCACAGCCACTGATCC

**Cloning 5'→3'**

attB1_PAX	GGGGACAAGTTTGTACAAAAAAGCAGGCTTTATGCTGGAAATGGAAAGAGTTGC
attB2_PAX	GGGGACCACTTTGTACAAGAAAGCTGGGTGAAAAAAGCTCAAAGTCTAGAAATTTACC
attB1r_prPAX	GGGGACTGCTTTTTGTACAACTTGTTTCATGCGATTTTAAACACCAAACAC
attB4_prPAX	GGGGACAAGTTTGTATAGAAAAGTTGTTTCGGACCAATTCATTCTTCAGAC
prCVP2	CGCAGAGCTTGAGTAAATCAAG
prCVP2	AACACAAGACAATGAAGCCATCATTGTTGCTTCTCTGCAAGTG
prCLE45	CATCAAAGATCTGATTGGTCACATC
prCLE45	AACACAAGACAATGAAGCCATCATTTCTGCTCTTAGGCAGACAAGA
mpd-2	ATGATGGCTTCATTGTCTTGTT
mpd-2	TGGTCACCAATTCACACGTGTTAGGTTTCGGACGCGGGGTG
attB1_PAX	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAATGCTGGAAATGGAAAGAGTTG
attB2_PAX with stop codon	GGGGACCACTTTGTACAAGAAAGCTGGGTATTAGAAAAAAGCTCAAAGTCTAGAAATTTACC
PAX_S596Dmut	[Phos]-GGGTTCCAACAAAGTCCATGGACCGTG
PAX_S596Amut	[Phos]-GGGTTCCAACAAATGCCATGGACCGTG



# Disruption of *TET2* promotes the therapeutic efficacy of CD19-targeted T cells

Joseph A. Fraietta<sup>1,2,3,4</sup>, Christopher L. Nobles<sup>5</sup>, Morgan A. Sammons<sup>6,10</sup>, Stefan Lundh<sup>1,2</sup>, Shannon A. Carty<sup>2,11</sup>, Tyler J. Reich<sup>1,2</sup>, Alexandria P. Cogdill<sup>1,2</sup>, Jennifer J. D. Morrisette<sup>3</sup>, Jamie E. DeNizio<sup>7,8</sup>, Shantanu Reddy<sup>5</sup>, Young Hwang<sup>5</sup>, Mercy Gohil<sup>1,2</sup>, Irina Kulikovskaya<sup>1,2</sup>, Farzana Nazimuddin<sup>1,2</sup>, Minnal Gupta<sup>1,2</sup>, Fang Chen<sup>1,2</sup>, John K. Everett<sup>5</sup>, Katherine A. Alexander<sup>6</sup>, Enrique Lin-Shiao<sup>6</sup>, Marvin H. Gee<sup>9</sup>, Xiaojun Liu<sup>1,2</sup>, Regina M. Young<sup>1,2</sup>, David Ambrose<sup>1,2</sup>, Yan Wang<sup>1,2</sup>, Jun Xu<sup>1,2</sup>, Martha S. Jordan<sup>2,3</sup>, Katherine T. Marcucci<sup>1,2</sup>, Bruce L. Levine<sup>1,2,3</sup>, K. Christopher Garcia<sup>9</sup>, Yangbing Zhao<sup>1,2</sup>, Michael Kalos<sup>1,2,3</sup>, David L. Porter<sup>1,2,7</sup>, Rahul M. Kohli<sup>5,7,8</sup>, Simon F. Lacey<sup>1,2,3</sup>, Shelley L. Berger<sup>6</sup>, Frederic D. Bushman<sup>5</sup>, Carl H. June<sup>1,2,3,4,\*</sup> & J. Joseph Melenhorst<sup>1,2,3,4,\*</sup>

**Cancer immunotherapy based on genetically redirecting T cells has been used successfully to treat B cell malignancies<sup>1–3</sup>. In this strategy, the T cell genome is modified by integration of viral vectors or transposons encoding chimaeric antigen receptors (CARs) that direct tumour cell killing. However, this approach is often limited by the extent of expansion and persistence of CAR T cells<sup>4,5</sup>. Here we report mechanistic insights from studies of a patient with chronic lymphocytic leukaemia treated with CAR T cells targeting the CD19 protein. Following infusion of CAR T cells, anti-tumour activity was evident in the peripheral blood, lymph nodes and bone marrow; this activity was accompanied by complete remission. Unexpectedly, at the peak of the response, 94% of CAR T cells originated from a single clone in which lentiviral vector-mediated insertion of the CAR transgene disrupted the methylcytosine dioxygenase *TET2* gene. Further analysis revealed a hypomorphic mutation in this patient's second *TET2* allele. *TET2*-disrupted CAR T cells exhibited an epigenetic profile consistent with altered T cell differentiation and, at the peak of expansion, displayed a central memory phenotype. Experimental knockdown of *TET2* recapitulated the potency-enhancing effect of *TET2* dysfunction in this patient's CAR T cells. These findings suggest that the progeny of a single CAR T cell induced leukaemia remission and that *TET2* modification may be useful for improving immunotherapies.**

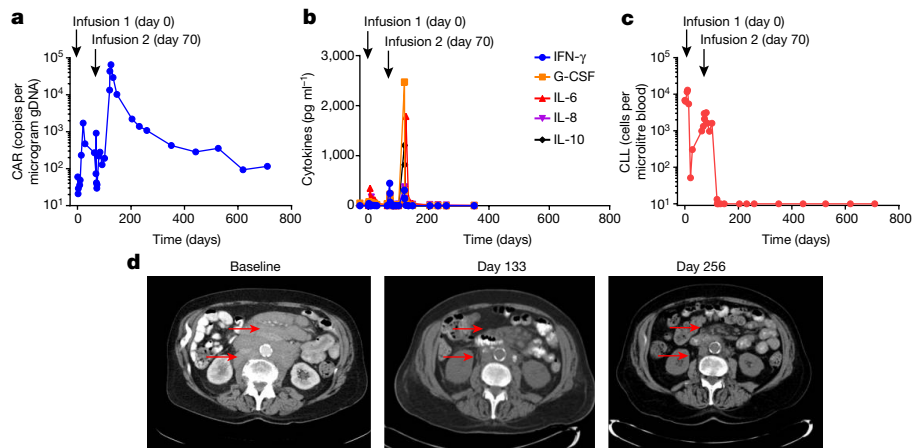
Here we describe an unusual case in which CAR T cell therapy was used to treat chronic lymphocytic leukaemia (CLL) that sheds light on the determinants of CAR-T cell efficacy and persistence. A seventy-eight-year-old man with advanced relapsed/refractory CLL (Patient-10, Supplementary Table 1) enrolled in a clinical trial for CD19 CAR T cell (CTL019) therapy (trial no. NCT01029366). He underwent two adoptive transfers of autologous CTL019 cells, approximately two months apart. Following the first infusion, he became persistently febrile and was diagnosed with cytokine release syndrome (CRS). Signs of CRS rapidly resolved following administration of interleukin (IL)-6 receptor-blocking therapy. Patient-10 continued to show progressive leukaemia six weeks after receiving his first dose of CAR T cells (Fig. 1a–c).

Because there was a concern that early blockade of IL-6-mediated signalling may have diminished the response to CAR T cell therapy, this patient was retreated with the remainder of his CAR T cells 70 days after the first dose (Supplementary Table 1). Infusions were again complicated by CRS, but this resolved without anti-IL-6 receptor-blocking

intervention. Evaluation of the patient's bone marrow one month later revealed extensive infiltration of CLL (Extended Data Fig. 1), and computed tomography (CT) scans showed minimal improvement in extensive adenopathy. Unexpectedly, two months after the second infusion, the expansion of CAR T cells peaked in the peripheral blood, followed by contraction (Fig. 1a). CTL019 cell outgrowth occurred mostly in the CD8<sup>+</sup> T cell compartment, which is typical in patients with CLL who respond to CAR T cell treatment (Extended Data Fig. 2a). Delayed CAR T cell expansion was accompanied by high-grade CRS and elevated circulating levels of interferon (IFN)- $\gamma$ , granulocyte-colony stimulating factor (G-CSF), IL-6, IL-8 and IL-10 (Fig. 1b). Coincident with the onset of high fever, rapid clearance of CLL was observed (Fig. 1c, d). Next-generation sequencing of rearrangement products at the immunoglobulin heavy chain (IGH) locus showed a 1-log reduction in tumour burden 51 days after the second infusion, with complete eradication of this tumour clone from the blood one month later (Supplementary Table 2). CT scans showed a marked improvement in mediastinal and axillary adenopathy (69% change; Fig. 1d). Patient-10 achieved a complete response with no evidence of CLL in his marrow (Extended Data Fig. 1, Supplementary Table 2) and resolution of all abnormal adenopathy six months later (Fig. 1d and data not shown). His most recent long-term follow-up evaluation (after more than 4.2 years) revealed the presence of CAR T cells in the peripheral blood, ongoing B cell aplasia (Extended Data Fig. 2b–e) and no evidence of circulating disease or marrow infiltration (Extended Data Fig. 1). Immune cell populations in the blood were normal in frequency, with no observed signs of lymphoproliferative abnormalities (Extended Data Fig. 2f and data not shown). The patient remains well in complete remission that has been sustained for more than five years at the time of this report.

Deep sequencing of the T cell receptor beta repertoire indicated that pre-infusion CD8<sup>+</sup> CTL019 cells and the peripheral blood CD8 T cell compartment one month after the second infusion were polyclonal, with multiple distinct TCRV $\beta$  clonotypes that were similar between the samples (Fig. 2a, Extended Data Fig. 3a). Approximately two months after the second infusion, TCRV $\beta$ 5.1 family usage exhibited a skewing of greater than 50%, with clonal dominance occurring in CD8<sup>+</sup> CTL019 cells (Fig. 2a, b). Subsequent analysis revealed that 94% of the CD8<sup>+</sup> CAR T cell repertoire consisted of a single clone that was not detected at the time of transfer or one month after the second infusion (Fig. 2c). The expansion of this clonal population of cells declined in

<sup>1</sup>Center for Cellular Immunotherapies, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>2</sup>Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>3</sup>Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>4</sup>Parker Institute for Cancer Immunotherapy, University of Pennsylvania, Philadelphia, PA, USA. <sup>5</sup>Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>6</sup>Department of Cell and Developmental Biology, Epigenetics Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>7</sup>Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>8</sup>Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>9</sup>Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. <sup>10</sup>Present address: Department of Biology, University at Albany, State University of New York, Albany, NY, USA. <sup>11</sup>Present address: Department of Internal Medicine and Rogel Cancer Center, University of Michigan, Ann Arbor, MI, USA. \*e-mail: [cjune@upenn.edu](mailto:cjune@upenn.edu); [mej@upenn.edu](mailto:mej@upenn.edu)



**Fig. 1 | Evaluation of clinical responses following adoptive transfer of CAR T cells in a patient with CLL.** **a**, In vivo expansion and persistence of CAR T cells. **b**, Longitudinal measurements of serum cytokines before and after CAR T cell infusions. **c**, Total number of circulating B-CLL

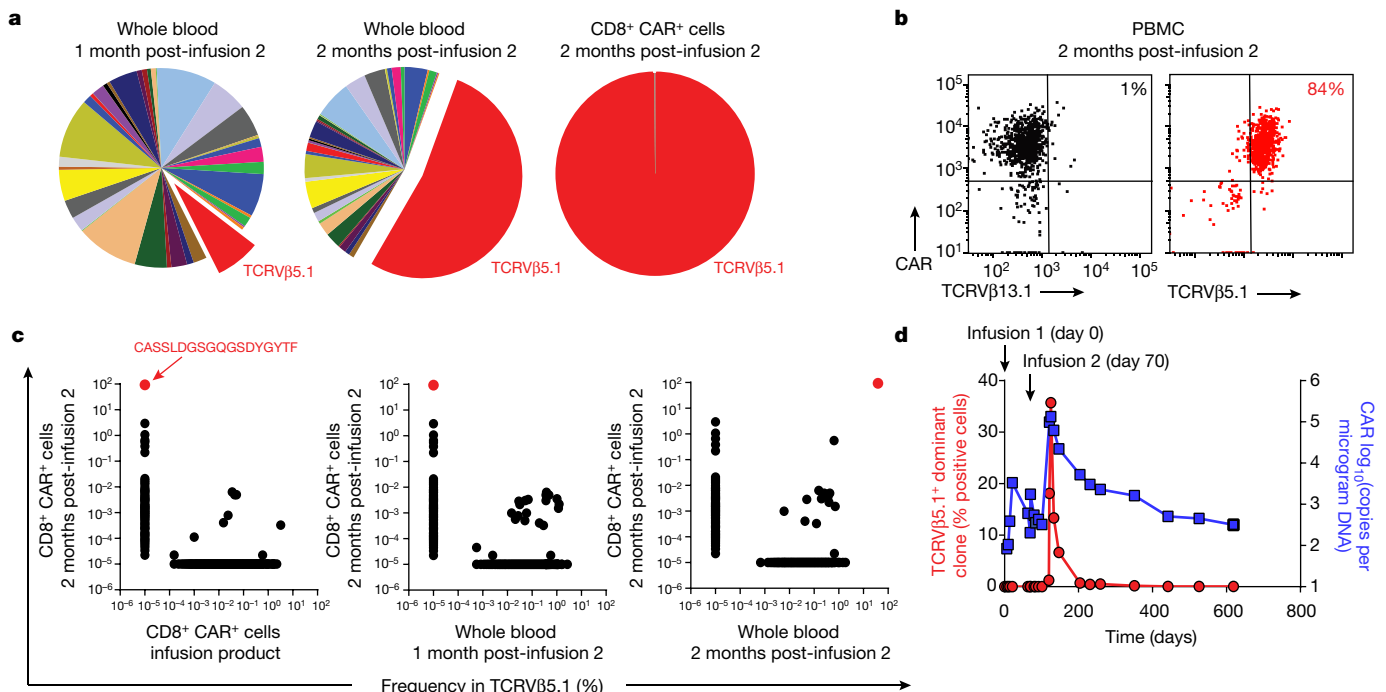
cells before and after CTL019 therapy. **d**, Sequential CT imaging of chemotherapy-refractory lymphadenopathy. Red arrows indicate masses that were progressively reduced following the second CAR T cell infusion.

line with CAR T cell decay kinetics (Fig. 2d). Thus, leukaemia was eliminated in this patient primarily by the progeny of a single CAR T cell that demonstrated massive in vivo expansion (approximately 29 population doublings in the peripheral blood).

Longitudinal analysis of blood samples from Patient-10 revealed a highly abundant cell clone with an integration site in intron 9 of *TET2*, which was expanded in CAR T cells at the peak of clinical activity and not at earlier time points (Fig. 3a). This large degree of T cell clonal dominance has not been observed in more than forty leukaemia patients treated with CD19-directed T cells at the University of Pennsylvania, as determined by lentiviral integration site analysis (Extended Data Fig. 3b–d and data not shown). In CLL and acute lymphoblastic leukaemia (ALL), accumulation of CAR T cells in vivo

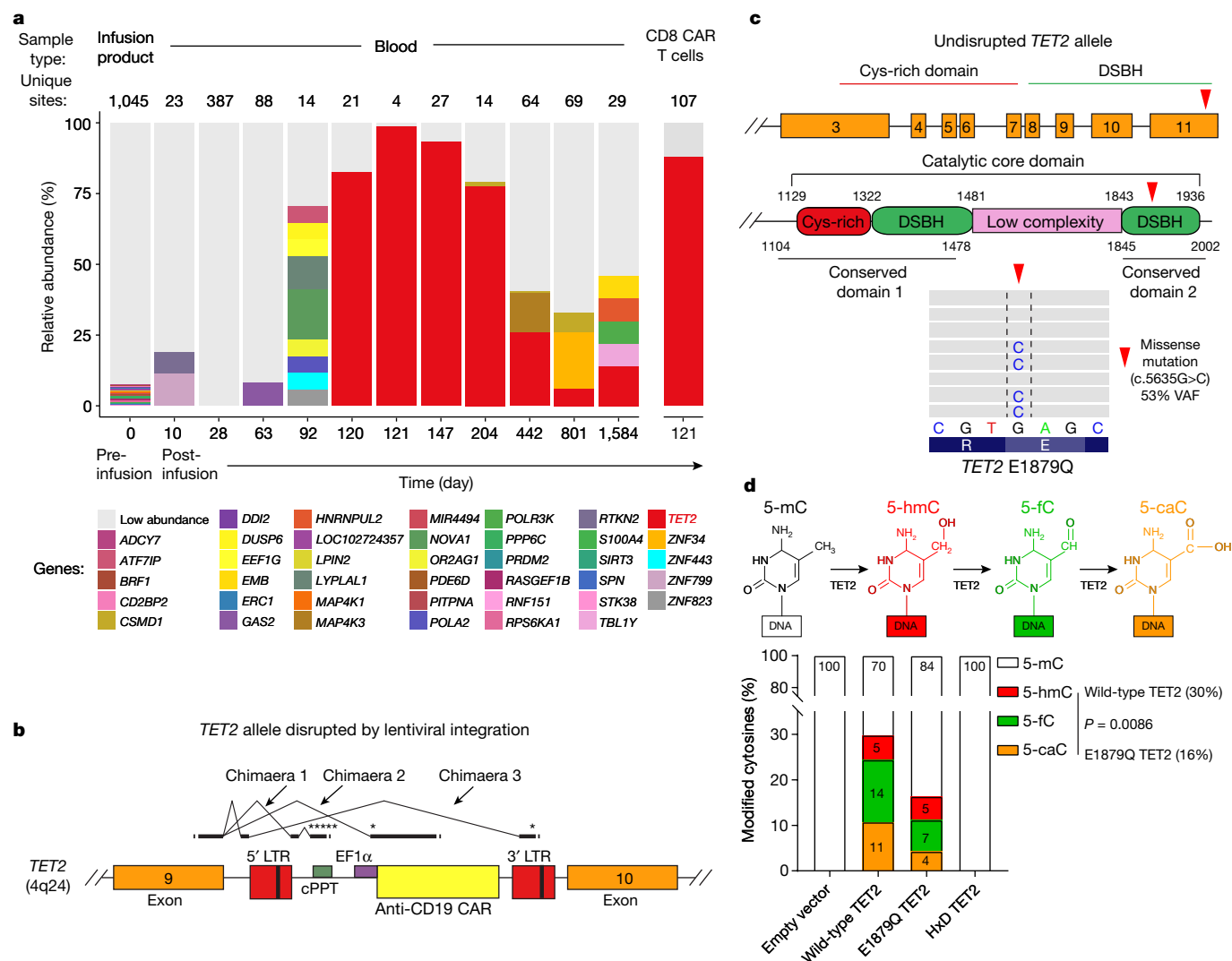
results from the expansion of a diverse polyclonal or pauciclonal repertoire within the transduced T cell population<sup>6</sup>. In Patient-10, cells bearing the *TET2* integration event were present in the blood at a relative abundance of 14% at 4.2 years after the infusion (Fig. 3a). The clonal population thus contracted, with no signs of insertional oncogenesis. Because *TET2* is a tumour suppressor gene, we are continuing to monitor this patient carefully.

*TET2* is a master regulator of blood cell formation. Haploinsufficiency or deletion of *TET2* is found in normal clonal haematopoiesis<sup>7</sup> as well as the initiation of lymphoma and leukaemia, including naturally arising and human T-lymphotropic virus type 1 (HTLV-1)-associated malignancies<sup>8–11</sup>. Although *TET2* inactivation may contribute to increased self-renewal of haematopoietic stem and progenitor cells, its



**Fig. 2 | Analysis of CAR T cell clonal expansion in a patient with CLL who had a delayed therapeutic response.** **a**, Frequency of TCRVβ gene segment usage in the blood of Patient-10 one month (left) and two months (middle) after the second CAR T cell infusion. TCRVβ clonotype frequencies in sorted CAR T cells at the peak of expansion are also shown (right). **b**, Flow cytometric proportions of TCRVβ5.1- versus TCRVβ13.1-positive (negative control) CAR T cells. **c**, Abundance of

different TCRVβ5.1 clones in CAR<sup>+</sup> T cells at the peak of activity relative to other time points. Red dots indicate the dominant TCRVβ5.1 clone (representative of two independent experiments). **d**, Expansion kinetics of the TCRVβ5.1<sup>+</sup> dominant clone following a second infusion of CAR T cells plotted in parallel with CTL019 levels. Percentages of positive cells were calculated from the results of a quantitative PCR assay designed to amplify clonotype-specific sequences.



**Fig. 3 | Investigation of CAR lentiviral integration sites and *TET2* dysfunction in Patient-10. a**, Longitudinal CAR T cell clonal abundance as marked by integration sites. Different colours (horizontal bars) indicate major cell clones. A key to the sites, named for the nearest gene, is shown below the graph (abundances below 3% binned as 'Low abundance').

**b**, Diagram of the vector at the *TET2* integration site locus illustrating splicing into the vector provirus to yield truncated transcripts. Asterisks denote ectopic stop codons. LTR, long terminal repeat; cPPT, polypurine tract; EF1 $\alpha$ , elongation factor 1- $\alpha$  promoter. **c**, Domain organization of *TET2* and location of a residue mutated in CAR<sup>+</sup> and CAR<sup>+</sup> T cells from Patient-10. Cysteine (Cys)-rich and catalytic double-stranded  $\beta$ -helix

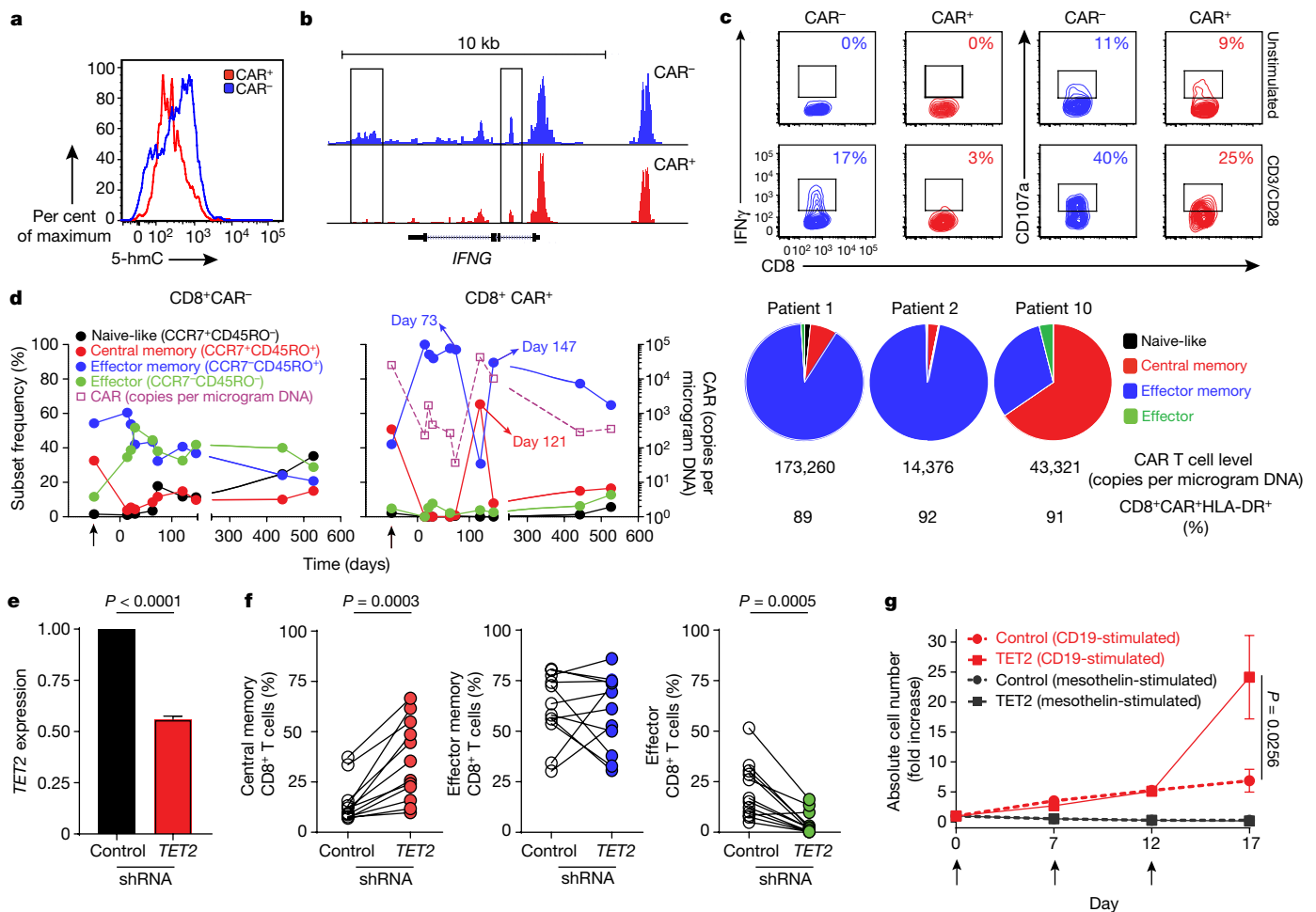
(DSBH) domains are also shown (red arrow denotes an amino acid change resulting from a missense mutation). Representative results of next generation sequencing appear beneath the structural diagrams. Each grey bar denotes a DNA fragment. A single-base G/C nucleotide substitution is highlighted by dashed lines above the consensus sequence. **d**, Diagram of the *TET2*-catalysed sequential oxidations of 5-mC to 5-hmC and to 5-fC and 5-caC (top). Genomic levels of 5-mC, 5-hmC, 5-fC, and 5-caC modifications produced by the E1879Q *TET2* mutant shown as per cent of total cytosine modifications. Percentages derived from the mean of three independent experiments are shown.  $P$  values were determined using a two-tailed, paired Student's  $t$ -test.

disruption alone infrequently leads to overt oncogenesis<sup>10,12</sup>. Analysis of polyadenylated *TET2* RNA populations in clonal CAR T cells at the peak of expansion in Patient-10 showed the appearance of new chimaeric RNAs that spliced from *TET2* exon 9 into the vector and terminated, truncating the encoded protein with premature stop codons (Fig. 3b, Extended Data Fig. 4a, b). Although it is possible that expression of the truncated fusion *TET2* protein may have a dominant-negative effect, it has been demonstrated that other *TET2* mutant proteins do not exhibit such characteristics<sup>13</sup>.

We next sequenced CAR<sup>+</sup> and CAR<sup>+</sup> T cells from this subject and examined genes involved in haematologic malignancies (Extended Data Fig. 4c). In both samples, a missense variant in *TET2* encoding amino acid 1879 was found in the catalytic domain, converting the wild-type residue, glutamic acid, to glutamine (Fig. 3c and data not shown). The c.5635C mutation was present in the allele of *TET2* without the integrated CAR transgene, as that allele contained the wild-type reference sequence (c.5635G; Extended Data Fig. 4d).

No other mutations were found in the panel of 67 additional genes analysed.

*TET2* encodes methylcytosine dioxygenase, an enzyme that catalyses the conversion of 5-methylcytosine (5-mC) into 5-hydroxymethylcytosine (5-hmC) and the rarer bases 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC), thereby mediating DNA demethylation<sup>14–19</sup> (Fig. 3d, top panel). Methylation at the C5 position of cytosine normally represses transcription, and therefore, demethylation is expected to activate gene expression. We interrogated the functional significance of the E1879Q mutation using plasmids encoding wild-type *TET2* or this *TET2* variant that were transfected into HEK293T cells (Extended Data Fig. 5a–c). Overexpression of wild-type or mutant *TET2* proteins was verified by western blotting (Extended Data Fig. 5b, c). Analysis of genomic DNA isolated from these cells using both dot blotting (Extended Data Fig. 5a, b) and liquid chromatography–tandem mass spectrometry (LC–MS/MS) (Extended Data Fig. 5d, e) revealed that E1879Q compromises the step-wise



**Fig. 4 | Effect of *TET2* alteration on the epigenetic landscape of CAR T cells.** **a**, Total 5-hmC levels in CAR<sup>+</sup> and CAR<sup>-</sup> T cells cultured from Patient-10. Histograms depict the intensity of intracellular 5-hmC staining. **b**, Genome browser views of ATAC enrichment at the *IFNG* locus of T cells from Patient-10. **c**, Frequencies of IFN $\gamma$ - and CD107a-expressing T cells expanded from Patient-10, unstimulated or stimulated with anti-CD3/CD28 antibodies. Insets indicate frequencies of gated cell populations. **d**, Longitudinal differentiation phenotypes of CAR<sup>-</sup> (left) and CAR<sup>+</sup> (middle) CD8<sup>+</sup> T cells from Patient-10. Black arrows denote pre-infusion CAR T cells. Differentiation phenotype at the peak of in vivo activity is shown in two patients with CLL who showed long-term complete responses (Patient-1 and Patient-2) compared to Patient-10 (right). Pie slices represent T cell subset frequencies. The CTL019 cell level

and frequencies of activated CAR T cells expressing HLA-DR (surface molecule indicative of T cell activation) at the peak of each patient's response are listed below the pie charts. **e**, *TET2* expression in healthy donor CD8<sup>+</sup> T cells transduced with a scrambled shRNA (control) or *TET2* shRNA. Error bars depict s.e.m. **f**, Frequencies of central memory (left), effector memory (middle) and effector (right) CD8<sup>+</sup> T cells following shRNA-mediated knockdown of *TET2* ( $n = 12$ ; pooled results from 4 independent experiments). **g**, Proliferation of healthy donor CTL019 cells ( $n = 8$ ; 3 independent experiments) in response to repetitive stimulation (denoted by black arrows) with K562 cells expressing CD19 or mesothelin (negative control). CAR T cells were transduced to express either a scrambled control or *TET2*-specific shRNA.  $P$  values were determined using a two-tailed, paired Student's  $t$ -test. All error bars depict s.e.m.

oxidation of 5-mC relative to wild-type *TET2*, with reductions in 5-fC and 5-caC occurring (Fig. 3d). Therefore, the clonal expansion of CAR T cells in Patient-10 was comprised of a compound heterozygous loss-of-function in *TET2* on one allele and a hypomorphic E1879Q variant on the other allele.

CAR<sup>+</sup>V $\beta$ 5.1<sup>+</sup> T cells from Patient-10 exhibited lower total levels of 5-hmC compared to their CAR<sup>-</sup>V $\beta$ 5.1<sup>-</sup> T cell counterparts (Fig. 4a). This was presumably the result of disruption of the wild-type *TET2* allele following lentiviral integration, as the *TET2* E1879Q variant can form 5-hmC (Fig. 3d, Extended Data Fig. 5a, b). To investigate the effects of *TET2* alteration on CAR T cell fate and function, we performed assay for transposase-accessible chromatin using sequencing (ATAC-seq), which monitors DNA accessibility (Supplementary Table 3). The global epigenetic changes between CAR<sup>+</sup> and CAR<sup>-</sup> T cells from Patient-10 were modest (Extended Data Fig. 6a). However, we found that genes with more accessible chromatin in CAR<sup>+</sup> compared to CAR<sup>-</sup> T cells were enriched in pathways that regulate the cell cycle and T cell receptor signalling (Extended Data Fig. 6b, Supplementary Tables 4, 5a). ATAC-seq peaks that were reduced or lost

in the setting of *TET2* biallelic dysfunction included those corresponding to several regulators of T cell effector differentiation or exhaustion such as *IFNG*, *NOTCH2*, *CD28*, *ICOS* and *PRDM1* (Fig. 4b, Extended Data Fig. 6c, Supplementary Tables 4, 5b). Furthermore, transcription factor motifs within sites that changed accessibility could have affected specific transcriptional circuits controlling CAR T cell fate and anti-tumour activity in this patient (Extended Data Fig. 6d, Supplementary Table 6a, b). Functional analysis of CAR<sup>+</sup> T cells with biallelically altered *TET2* cultured from this patient showed a diminished capacity to express IFN $\gamma$  and CD107a (a degranulation marker) when activated (Fig. 4c), consistent with a less differentiated state. Thus, lentiviral integration into *TET2* together with a hypomorphic mutation on the second allele reprogrammed the epigenetic landscape of CAR T cells in a manner that was consistent with altered T lymphocyte differentiation.

We next analysed the differentiation state of ex vivo CTL019 cells from Patient-10 and compared it to those of CAR T cells from six other patients who responded to this therapy (Extended Data Fig. 7), including two subjects with CLL (Patient-1 and Patient-2) who had long-term durable remissions (of more than 6 years) and did not have



*TET2* integrations (data not shown). At the peak of in vivo expansion and activation marker expression, 65% of the CAR T cells in Patient-10 had a central memory phenotype (Fig. 4d, Extended Data Fig. 7a) which differed from other responders whose repertoires were dominated by CD8<sup>+</sup> effector memory and effector CTL019 cells at the height of the response (Fig. 4d, Extended Data Fig. 7b). Experimental knockdown of *TET2* (Fig. 4e) recapitulated the effect of its dysfunction in Patient-10 on the differentiation state of both total and CAR<sup>+</sup>CD8<sup>+</sup> as well as CD4<sup>+</sup> primary T cells (Fig. 4f, Extended Data Fig. 8a, b), implicating *TET2* as an epigenetic modulator of human T lymphocyte fate. *TET2*-mediated regulation of CD8<sup>+</sup> T cell differentiation may not occur at the transcriptional level, as we did not observe differential *TET2* mRNA expression between naive and memory subsets (Extended Data Fig. 8c).

To investigate the effects of *TET2* inhibition on CAR T cell function, we performed an in vitro serial re-stimulation assay. Repeated stimulation with CD19-expressing tumour cells enabled *TET2* knockdown CAR T cells to continue to expand in an antigen-dependent manner, whereas re-stimulation of CAR T cells with unaltered *TET2* resulted in arrest of culture growth (Fig. 4g) without affecting viability (Extended Data Fig. 8d). We next examined cytokine production following acute (Extended Data Fig. 9a) and chronic (Extended Data Fig. 9b, c) antigen stimulation. Consistent with our analysis of CD8<sup>+</sup>CAR<sup>+</sup> T cells in Patient-10, IFN $\gamma$  production following activation of CD3 and CD28 was diminished in CD8<sup>+</sup> as well as CD4<sup>+</sup> T cells with reduced *TET2* levels (Extended Data Fig. 9a). A similar decrease was observed for TNF $\alpha$  generation (Extended Data Fig. 9a). By contrast, acute production of both TNF $\alpha$  and IL-2 by CD4<sup>+</sup> T cells was increased upon CAR-specific stimulation (Extended Data Fig. 9a). While repeated exposure of bulk CTL019 cells to CD19-expressing tumours also led to decreased IFN $\gamma$  elaboration (Extended Data Fig. 9b, c), *TET2* inhibition resulted in the sustained production of various other cytokines following multiple rounds of stimulation (Extended Data Fig. 9b). Thus, *TET2* may control human T cell subset-specific cytokine production in an antigen-receptor-dependent and/or co-stimulatory signal-dependent fashion.

On the basis of our evaluation of CAR<sup>+</sup> T cells expanded from Patient-10, we predicted that knockdown of *TET2* would decrease effector molecule expression. Unexpectedly, CAR-specific stimulation, but not CD3 and CD28 stimulation, increased the expression of CD107a (Extended Data Fig. 10a). This may be attributed, at least in part, to enhanced cytolytic capacity mediated by 4-1BB over CD28 co-stimulation due to NKG2D upregulation<sup>20</sup>. Because CD8<sup>+</sup> T cell differentiation is accompanied by decreased methylation and upregulated gene expression at effector gene loci, including *GZMB* (encoding granzyme B) and *IFNG*<sup>21</sup>, we subsequently investigated whether *TET2* inhibition influences critical components of the cytotoxic machinery. In contrast to IFN $\gamma$ , *TET2* reduction in CD8<sup>+</sup> CAR<sup>+</sup> T cells increased granzyme B and perforin expression (Extended Data Fig. 10b). These changes were associated with the heightened cytotoxic activity of *TET2* knockdown CAR T cells (Extended Data Fig. 10c).

The above findings suggest that *TET2* dysfunction may produce potent CAR T cells with properties of short-lived memory cells that can expand and elicit effector responses, as well as long-lived, persistent memory cells. We thus examined additional effector or memory markers in CD8<sup>+</sup>CAR<sup>+</sup> and CAR<sup>-</sup> T cells using post-infusion samples from Patient-10 and other long-term responding patients with CLL. At the height of the response, tumour-reactive CAR<sup>+</sup> T cells from Patient-10 possessed higher levels of granzyme B (Extended Data Fig. 11a) and eomesodermin (Eomes; a transcription factor involved in the formation and maintenance of the CD8<sup>+</sup> memory T cell pool; Extended Data Fig. 11b) compared to matched CAR<sup>-</sup> T cells, unlike the other complete responders. All clinically active CD8<sup>+</sup> CAR<sup>+</sup> T cells in these patients expressed CD27, a co-stimulatory receptor involved in the generation of T cell memory (Extended Data Fig. 11b). The frequency of CTL019 cells expressing KLRG1, a marker of T lymphocyte senescence that is known to be regulated by DNA methylation<sup>21</sup>, was significantly lower

on CAR<sup>+</sup> T cells from Patient-10 than on cells from other patients (Extended Data Fig. 11b). A high frequency of Ki-67-positive CAR<sup>+</sup> T cells was observed at the peak of in vivo expansion in Patient-10 (Extended Data Fig. 11a), further suggesting that *TET2* is required for CAR-specific CD8<sup>+</sup> T cell proliferation. These observations collectively support the idea that *TET2* dysfunction promotes the development of human memory CAR T cells that can elicit effective anti-tumour responses. The remission observed in Patient-10 was likely to result from the marked increase in the number of CTL019 cells with *TET2* dysfunction, despite the reduction in certain effector functions.

In summary, profound clonal expansion of a single CAR-transduced T cell with biallelic *TET2* dysfunction transformed a non-curative response into a deep molecular remission in a seventy-eight-year-old patient with CLL. Characterization of this T lymphocyte population revealed that changes to the epigenetic environment altered the differentiation state and proliferative capacity of the cells and translated into a considerable therapeutic effect. Although our initial studies were based on an extensive analysis of one subject, recapitulation of the effect of *TET2* dysfunction on CAR T cell fate and anti-tumour activity in relevant culture systems involving primary human T-lymphocytes supports the discovery of a modifiable epigenetic pathway that can shape the immune response. Thus, targeting the epigenome may improve the efficacy and persistence of CAR T cells. Finally, our results indicate that the progeny of a single CAR T cell are sufficient to mediate potent anti-tumour effects in advanced leukaemia, a finding that has substantial clinical implications regarding the delivery of bespoke cellular therapies.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0178-z>.

Received: 19 February 2017; Accepted: 27 April 2018;

Published online: 30 May 2018

- Porter, D. L., Levine, B. L., Kalos, M., Bagg, A. & June, C. H. Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *N. Engl. J. Med.* **365**, 725–733 (2011).
- Schuster, S. J. et al. Chimeric antigen receptor T cells in refractory B cell lymphomas. *N. Engl. J. Med.* **377**, 2545–2554 (2017).
- Maude, S. L. et al. Tisagenlecleucel in children and young adults with B cell lymphoblastic leukemia. *N. Engl. J. Med.* **378**, 439–448 (2018).
- Porter, D. L. et al. Chimeric antigen receptor T cells persist and induce sustained remissions in relapsed refractory chronic lymphocytic leukemia. *Sci. Transl. Med.* **7**, 303ra139 (2015).
- Savoldo, B. et al. CD28 costimulation improves expansion and persistence of chimeric antigen receptor-modified T cells in lymphoma patients. *J. Clin. Invest.* **121**, 1822–1826 (2011).
- Turtle, C. J. et al. CD19 CAR-T cells of defined CD4<sup>+</sup>:CD8<sup>+</sup> composition in adult B cell ALL patients. *J. Clin. Invest.* **126**, 2123–2138 (2016).
- Busque, L. et al. Recurrent somatic *TET2* mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet.* **44**, 1179–1181 (2012).
- Tefferi, A. et al. Detection of mutant *TET2* in myeloid malignancies other than myeloproliferative neoplasms: CMML, MDS, MDS/MPN and AML. *Leukemia* **23**, 1343–1345 (2009).
- Delhommeau, F. et al. Mutation in *TET2* in myeloid cancers. *N. Engl. J. Med.* **360**, 2289–2301 (2009).
- Quivoron, C. et al. *TET2* inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25–38 (2011).
- Yeh, C. H. et al. Mutation of epigenetic regulators *TET2* and *MLL3* in patients with HTLV-I-induced acute adult T cell leukemia. *Mol. Cancer* **15**, 15 (2016).
- Zang, S. et al. Mutations in 5-methylcytosine oxidase *TET2* and RhoA cooperatively disrupt T cell homeostasis. *J. Clin. Invest.* **127**, 2998–3012 (2017).
- Aslanyan, M. G. et al. Clinical and biological impact of *TET2* mutations and expression in younger adult AML patients treated within the EORTC/GIMEMA AML-12 clinical trial. *Ann. Hematol.* **93**, 1401–1412 (2014).
- Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner *TET1*. *Science* **324**, 930–935 (2009).
- Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
- Ito, S. et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).
- Pfaffeneder, T. et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem.* **50**, 7008–7012 (2011).

18. He, Y. F. et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
19. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
20. Zhang, H. et al. 4-1BB is superior to CD28 costimulation for generating CD8+ cytotoxic lymphocytes for adoptive immunotherapy. *J. Immunol.* **179**, 4910–4918 (2007).
21. Scharer, C. D., Barwick, B. G., Youngblood, B. A., Ahmed, R. & Boss, J. M. Global DNA methylation remodeling accompanies CD8 T cell effector function. *J. Immunol.* **191**, 3419–3429 (2013).

**Acknowledgements** L. Tian, V. Gonzalez, N. Kengle, J. Scholler, Y. Wu, A. Bagg, C. Pletcher, B. Carreno, A. Bigdeli and A. Chew are acknowledged for research support. D. Campana and C. Imai provided the CD19-directed CAR under material transfer agreements. B. Jena and L. Cooper provided the CAR anti-idiotypic antibody. The OSU-CLL cell line was a kind gift from J. C. Byrd. This work was supported by funding from NCI T32CA009140 (J.A.F.), P01CA214278 (C.H.J.), AI104400, AI 082020, AI045008, AI117950 (F.D.B.), NIAID K08AI101008 (S.A.C.), NIGMS R01GM118501 (R.M.K.), R01CA165206 (D.L.P. and C.H.J.), a Stand Up to Cancer Phillip A. Sharp Innovation in Collaboration Award (S.L.B. and C.H.J.) and Novartis.

**Reviewer information** *Nature* thanks M. Maus and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** J.A.F., C.L.N., M.A.S., S.A.C., J.J.D.M., R.M.Y., M.S.J., K.T.M., B.L.L., K.C.G., Y.Z., M.K., D.L.P., R.M.K., S.F.L., S.L.B., F.D.B., C.H.J. and J.J.M. designed experiments. J.A.F., C.L.N., M.A.S., S.L., S.A.C., T.R., A.P.C., J.J.D.M., J.E.D., S.R., Y.H., M.G., I.K., F.N., M.G., F.C., J.K.E., K.A.A., E.L.-S., M.H.G., X.L., D.A., Y.W. and J.X. performed experiments and/or analysed data. J.A.F., F.D.B., C.H.J. and J.J.M. wrote and edited the manuscript, with all authors providing feedback.

**Competing interests** J.A.F., C.L.N., R.M.Y., B.L.L., M.K., D.L.P., S.F.L., F.D.B., C.H.J. and J.J.M. hold patents related to CTL019 cell therapy. These authors declare no additional competing interests. The remaining authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0178-z>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0178-z>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to C.H.J. or J.J.M.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

**Patient samples.** Patients were enrolled in institutional review board (IRB)-approved clinical protocol: 'Genetically engineered lymphocyte therapy in treating patients with B cell leukemia or lymphoma that is resistant or refractory to chemotherapy' (ClinicalTrials.gov number: NCT01029366) which was designed to evaluate the safety and efficacy of the adoptive transfer of autologous T cells expressing CD19 chimaeric antigen receptors (CAR19) that incorporate TCR- $\zeta$  and 4-1BB costimulatory domains (CTL019). Participants provided written informed consent in accordance with the Declaration of Helsinki and the International Conference on Harmonization Guidelines for Good Clinical Practice. All regulations were followed according to this United States (US) Food and Drug Administration (FDA)-approved clinical protocol. The current study is a secondary investigation using patient samples collected from existing clinical trials. Therefore, the sample sizes in this report were determined by the original clinical trial designs and sample availability; no additional inclusion/exclusion criteria were applied.

**Cell lines.** NALM-6, K562 and HEK293T cell lines were originally obtained from the American Type Culture Collection (ATCC). OSU-CLL cells<sup>22</sup> were obtained from Ohio State University. Cells were expanded in RPMI medium containing 10% fetal bovine serum (FBS), penicillin and streptomycin at a low passage and tested for mycoplasma using the MycoAlert detection kit as per the manufacturer's (Lonza) instructions. Authentication of cell lines was performed by the University of Arizona (USA) Genetics Core based on criteria established by the International Cell Line Authentication Committee. Short tandem repeat (STR) profiling revealed that these cell lines were well above the 80% match threshold. NALM-6 and OSU-CLL cells were engineered to constitutively express click beetle green (CBG) luciferase/enhanced GFP (eGFP). K562 cells were transduced with a lentiviral vector encoding human CD19 (K562-CD19), and negative control cells were K562 cells expressing mesothelin (K562-mesothelin)<sup>23</sup>. Following transgene introduction, cells were sorted on a FACS Aria (BD) to obtain a >99% pure population. Mycoplasma and authentication testing were routinely performed before and after molecular engineering.

**CAR T cell manufacturing and correlative studies.** Peripheral blood T cells for CTL019 cell manufacturing were obtained by leukapheresis as previously described<sup>1,4</sup>. The processing, flow cytometric evaluation, quantification of cytokines and quantitative PCR analyses (that is, detection of CAR19 and the TCRV $\beta$ 5.1<sup>+</sup> dominant clone) on pre- and post-CTL019 infusion samples were conducted as previously reported<sup>24</sup>. Next-generation sequencing of immunoglobulin heavy chain (IGH) rearrangements was carried out on DNA isolated from blood and marrow samples. In brief, primers specific for the variable and joining gene segments of the third complementarity-determining region of the IGH were used for amplification and deep sequencing to identify the leukaemic clone relative to baseline samples (Adaptive Biotechnologies). The frequency of the leukaemic clone in each sample was calculated using the number of total and unique productive reads.

Population doublings of clonal CAR T cells in the blood (assuming a 5-litre volume of peripheral blood) were calculated using the equation  $A_t = A_0 2^n$ , in which  $n$  is the number of population doublings,  $A_0$  is the input number of cells (assuming a single *TET2*-disrupted cell), and  $A_t$  is the number of CAR T cells at day 121 (peak CAR T cell expansion; 95.445 CD3<sup>+</sup> CD8<sup>+</sup> CAR<sup>+</sup> cells per microlitre whole blood).

These correlative assays were carried out at time points defined by the clinical protocol in parallel with disease response evaluations. This clinical trial was a single-treatment study; comparisons between patients in the current study were defined by the observed clinical responses. Investigators were blinded to clinical responses as correlative assays were conducted using de-identified subject samples.

**Flow cytometry.** Routine assessments of CAR T cell expansion and persistence as well as measurement of B-CLL burden in the blood and marrow were conducted according to our previously published methods using a six-parameter Accuri C6 flow cytometer (BD)<sup>1,4</sup>. T cell immunophenotyping was performed by surface staining with flow cytometry antibodies immediately following pre-incubation with Aqua Blue dead cell exclusion dye (Invitrogen). The Alexa Fluor 647-conjugated monoclonal antibody that was used to detect the CAR molecule has previously been described<sup>25</sup>. Commercially available flow cytometry antibodies against the following antigens used in the study are as follows: CD3 allophycocyanin (APC) H7, CCR7 PE/CF594, CD107a APC, (BD Biosciences); CD45RO brilliant violet (BV)570, CD8 BV650, CD4 BV785, perforin BV421, Ki-67 Alexa Fluor (AF)700, TNF $\alpha$  BV605 (Biolegend); TCRV $\beta$ 5.1 APC, IFN $\gamma$  PE, IL-2 PerCP-eFluor 710, Eomes FITC (eBioscience); Granzyme B PE/cyanine5.5 (Invitrogen). For intracellular staining, cells were fixed and permeabilized using the Foxp3 Fixation/Permeabilization Kit (eBioscience) or the Cytofix/Cytoperm Kit (BD Biosciences). The GolgiStop protein transport inhibitor containing monensin and GolgiPlug protein transport inhibitor containing brefeldin A (BD Biosciences) were used when staining for intracellular cytokine production. All flow cytometry reagents were titrated before use. Samples were acquired on an LSRFortessa (BD) and data were analysed using FlowJo software (TreeStar).

**TCRV $\beta$  deep sequencing.** Genomic DNA from pre-infusion T cells, peripheral blood samples or sorted post-infusion T cells was isolated using the DNeasy Blood and Tissue Kit (Qiagen). TCRV $\beta$  deep sequencing was carried out by immunoSEQ (Adaptive Biotechnologies). Only productive TCR rearrangements were used in the assessment of TCR clonotype frequencies.

**Integration site analysis.** Vector integration sites were detected from genomic DNA as described previously<sup>26–29</sup>. Genomic sequences were aligned to the human genome by BLAT (hg38, version 35, >95% identity) and statistical methods for analysing integration site distributions were carried out as previously described<sup>30</sup>. The SonicAbundance method was used to infer the abundance of cell clones from integration site data<sup>28</sup>. All samples were analysed independently in quadruplicate to suppress founder effects in the PCR and stochastics of sampling.

**Detection of *TET2* chimaeric transcripts.** RNA was isolated from cells and used as template with the One-Step RT-PCR Kit (Qiagen). Primers were designed to target the exon 9 and 10 boundaries of *TET2*, flanking the vector integration site and sequences internal to the anti-CD19BB $\zeta$  CAR lentiviral vector. These included various regions of the vector sequence (Extended Data Fig. 4a). Reactions were carried out as per the manufacturer's specifications. Thermocycling temperatures and time for reverse transcription as well as PCR activation were conducted using the following cycling conditions: 30 s at 94°C for melting, 30 s at 57°C for primer annealing and 1.5 min at 72°C for primer extension (35 cycles). A final extension at 72°C was held for 10 min for each sample. PCR products were visualized on ethidium bromide agarose gels (1.5% by weight) via electrophoresis and ultraviolet imaging.

**Next-generation sequencing of post-infusion CAR T cell samples.** CAR<sup>+</sup> and CAR<sup>−</sup> CD8<sup>+</sup> T cells were purified from post-infusion PBMC samples corresponding to the peak of in vivo expansion in Patient-10. T cells were sorted using a FACS Aria (BD) and genomic DNA was isolated from these lymphocytes as described above. A custom targeted next-generation sequencing panel of 68 genes associated with haematologic malignancies was then used (TruSeq Custom Amplicon, Illumina), and sequencing was carried out on an Illumina MiSeq (Illumina). A minimal mean depth of 2,110 reads was achieved for the specimens sequenced, with the assay and bioinformatics performed as previously described<sup>31</sup>. The data presented are based on human reference sequence UCSC build hg 19 (NCBI build 37.1).

**Identifying the *TET2* allele hosting vector integration.** A PCR assay was developed to amplify the region of DNA (approximately 4 kb) between the vector integration and the locus of the c.5635G>C mutation. Primers were designed to anneal to the vector sequence (MKL-3: 5'-CTTAAGCCTCAATAAGCTTGCCTTGAG-3') and multiple locations downstream of the mutation, chr4:105,276,145 (50 bp: 5' GCTGGTAAAGACGAGGGAGATCCTG-3', 99 bp: 5'-GGCTTCCCAAAGAGCCAAGCCATG-3', 120 bp: 5'-CACGGGCTTTTTCAGCCATTTTGGC-3'). Genomic DNA samples from sorted CAR<sup>+</sup> and CAR<sup>−</sup> CD8<sup>+</sup> T cells corresponding to the peak of clonal expansion in Patient-10 were selected for amplification. PCR reactions were carried out with Long Amp Taq polymerase (New England Biolabs) and 100–400 ng of DNA isolated from samples, according to the manufacturer's recommendations. Amplification was conducted as follows: 94°C for 30 s, 30 cycles of (94°C for 30 s, 60°C for 30 s, and 65°C for 3 min 20 s), and a final extension of 65°C for 10 min. Amplified products were separated by electrophoresis on a 1% ethidium bromide agarose gel and prominent bands of 4 kb in size were isolated using the QIAquick Gel Extraction Kit (Qiagen). Isolated bands were ligated into pCR2.1 and cloned into TOP-10 chemically competent cells using the TOPO TA Cloning Kit (Invitrogen). Purified plasmids were sequenced with M13 forward and reverse primers using standard Sanger technology. Sequencing results were aligned to the vector sequence and reference genome.

**Characterization of the *TET2* E1879Q mutation.** The previously characterized and crystalized human *TET2*-CS variant (1129–1936  $\Delta$ 1481–1843) with an N-terminal FLAG-tag was expressed using a pLEXm expression vector<sup>32,33</sup>. The E1879Q mutation or mutation of the catalytic H1382Y and D1384A (HxD mutant) were generated by standard means. HEK293T cells were cultured in DMEM with GlutaMAX (Thermo Fisher Scientific) and 10% FBS (Sigma). Cells were transfected with wild-type (WT), mutant hTET2-CS or an empty pLEXm vector control using Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's protocol. Medium was changed 24 h after transfection, and cells were collected by trypsinization 48 h after transfection and resuspended in phosphate-buffered saline. Genomic DNA was isolated from four-fifths of the cells using the DNeasy Blood and Tissue Kit (Qiagen) and the remaining one-fifth of the cells were lysed using CytoBuster Protein Extraction Reagent (EMD Millipore) for western blot analysis.

DNA blots for cytosine modifications were carried out according to established protocols<sup>34</sup>. Purified DNA from HEK293T cells was diluted to 15 ng/ $\mu$ l in Tris-EDTA (TE) buffer, pH 8.0 for twofold dilutions of each sample. One-quarter volume of 2 M NaOH–50 mM EDTA was added to each sample. The DNA was



denatured for 10 min at 95 °C and transferred quickly to ice, followed by the addition of 1:1 ice cold 2 M ammonium acetate. Polyvinylidene difluoride (PVDF) membranes were cut to size, wet with MeOH and equilibrated in TE buffer and then assembled into the PR 648 Slot Blot Manifold (GE Healthcare Life Sciences). Each well was washed with 400 µl TE drawn through with gentle vacuum, and genomic DNA was loaded, followed by another TE wash. Membranes were blocked for 2 h in 5% milk-TBST, washed thrice with TBST, and blotted at 4 °C overnight with primary antibodies against each modified cytosine (1:5,000 mouse anti-5-mC (Abcam); 1:10,000 rabbit anti-5-hmC (Active Motif); 1:5,000 rabbit anti-5-fC (Active Motif); 1:10,000 rabbit anti-5-caC (Active Motif)). Blots were then washed, incubated with a 1:2,000 dilution of a secondary horse anti-mouse-horseradish peroxidase (HRP; Cell Signaling Technology) or 1:5,000 goat anti-rabbit-HRP (Santa Cruz Biotechnology) for 2 h, washed and imaged using the Immobilon Western Chemiluminescent HRP Substrate (Millipore) and the Amersham Imager 600 (GE Healthcare Life Sciences).

For protein detection, clarified cell lysates were run on 8% sodium dodecyl sulphate polyacrylamide (SDS–PAGE) gels. Gels were transferred together onto a PVDF membrane using the iBlot 2 Gel Transfer Device (Thermo Fisher Scientific). Membranes were blocked for 2 h at room temperature with 5% (w/v) milk in Tris-buffered saline with 0.1% (v/v) Tween-20 (TBST), washed three times with TBST and blotted either with primary 1:10,000 anti-FLAG M2 (Sigma) or 1:1,000 anti-Hsp90α/β (Santa Cruz Biotechnology) antibodies at 4 °C overnight. Following incubation, membranes were washed and blotted with a 1:5,000 dilution of secondary goat anti-mouse-HRP antibodies (Santa Cruz Biotechnology) for 2 h, washed and imaged with Immobilon Western Chemiluminescent HRP Substrate (Millipore) on an Amersham Imager 600 (GE Healthcare Life Sciences).

For LC–MS/MS, 1–2 µg of genomic DNA from each sample was degraded to component nucleosides with 1 U DNA Degradase Plus (Zymo Research Corporation) at 37 °C overnight. The nucleoside mixture was diluted tenfold into 0.1% formic acid, and injected onto an Agilent 1200 Series HPLC with a 5 µm, 2.1 × 250 mm Supelcosil LC-18-S analytical column (Sigma) equilibrated to 45 °C in buffer A (5 mM ammonium formate, pH 4.0). The nucleosides were separated in a gradient of 0–15% buffer B (4 mM ammonium formate, pH 4.0, 20% (v/v) methanol) over 8 min at a flow rate of 0.5 ml per minute. Tandem MS was performed by positive ion mode ESI on a 6460 triple-quadrupole mass spectrometer (Agilent) with a gas temperature of 250 °C, a gas flow of 12 l/min, a nebulizer pressure of 35 psi, a sheath gas temperature of 300 °C, a sheath gas flow of 11 l/min, a capillary voltage of 3,500 V, a fragmentor voltage of 70 V and a delta EMV of +1,000 V. Collision energies were optimized to 10 V for 5-mC and 5-fC; 15 V for 5-caC; and 25 V for 5-hmC. Multiple reaction monitoring (MRM) mass transitions were 5-mC 242.11 → 126.066 *m/z*; 5-hmC 258.11 → 124.051; 5-fC 256.09 → 140.046; 5-caC 272.09 → 156.041; and T 243.10 → 127.050. Standard curves were generated using standard nucleosides (Berry & Associates) ranging from 2.5 µM to 610 pM (12.5 pmol to 3 fmol total). Digested oligonucleotides containing equimolar amounts of each modified cytosine were used as quality control samples. The sample peak areas were fit to the standard curve, as adjusted by the quality control samples to determine the amount of each modified cytosine in the genomic DNA sample. Amounts are expressed as the percentage of total cytosine modifications.

**Measurement of total 5-hydroxymethylcytosine levels.** CD8<sup>+</sup> T cells were purified from post-infusion PBMC samples using the EasySep Human CD8<sup>+</sup> T cell Immunomagnetic Negative Selection Kit (StemCell Technologies) and expanded ex vivo using a previously reported rapid expansion protocol<sup>35</sup>. Following culture, CD8<sup>+</sup> CAR<sup>+</sup>TCRVβ5.1<sup>+</sup> and CD8<sup>+</sup> CAR<sup>−</sup>TCRVβ5.1<sup>−</sup> T cells were sorted on a FACSaria (BD). Cells were permeabilized and treated with 300 µg/ml DNase I for 60 min at 37 °C. After washing, samples were incubated with an anti-5-hmC monoclonal antibody or an isotype control for 30 min, followed by staining with an Alexa Fluor 647-conjugated secondary antibody, as previously described<sup>36</sup>. Cells were immediately acquired on an LSRFortessa (BD).

**Global chromatin profiling by ATAC-seq.** Following culture, CD8<sup>+</sup> CAR<sup>+</sup>TCRVβ5.1<sup>+</sup> and CD8<sup>+</sup> CAR<sup>−</sup>TCRVβ5.1<sup>−</sup> T cells were sorted on a FACSaria (BD). ATAC-seq was carried out as previously described<sup>37,38</sup>. Two replicates were performed for each ex vivo expanded CD8<sup>+</sup> CAR<sup>+</sup>TCRVβ5.1<sup>+</sup> and CD8<sup>+</sup> CAR<sup>−</sup>TCRVβ5.1<sup>−</sup> T cell culture. In brief, nuclei were isolated from 200,000 sorted CD8<sup>+</sup> T cells for each replicate, followed by the transposition reaction in the presence of Tn5 transposase (Illumina) for 45 min at 37 °C. Purification of transposed DNA was subsequently completed with the MinElute Kit (Qiagen) and fragments were barcoded with dual indexes (Illumina Nextera). Paired-end sequencing (2 × 75-bp reads) was carried out using the Illumina NextSeq 500. Raw sequencing data were processed and aligned to the GRC37/hg19 reference genome using Bowtie 2 and regions of significant enrichment were identified using MACS v1.4.2. Merged peak lists were created using BedTools. ATAC sequencing tag enrichment and DNA motif analysis across the merged peak list were carried out using HOMER (<http://homer.salk.edu>). Gene Ontology pathway analysis was

performed with Metascape (<http://metascape.org>). Only high-confidence peaks were used for gene ontology and DNA motif analyses. For these evaluations, peaks with an enrichment score less than 5 were filtered out as previously established<sup>39</sup>.

**CAR T cell differentiation and expansion potency assays.** Bulk primary human T cells were activated with paramagnetic beads coated with anti-CD3 and anti-CD28 monoclonal antibodies as previously described<sup>40</sup> and transduced with lentiviral vectors encoding the CD19Bβ CAR and shRNA hairpin sequences targeting *TET2* or a scrambled control with GFP co-expression (Cellecta). Knockdown efficiency in CTL019 cells following shRNA transduction was determined by real-time quantitative PCR with Taqman gene expression assays (Applied Biosystems) for *TET2* (assay Hs00325999\_m1) and GAPDH (assay Hs03929097\_g1), which served as a loading and normalization control. Following 14 days of culture, the differentiation phenotype of these cells was determined by flow cytometry. GFP<sup>+</sup>CAR<sup>+</sup> T cells were sorted on a FACSaria (BD) and combined 1:1 with irradiated K562 cells engineered to express CD19<sup>4</sup> or mesothelin as a negative control. CTL019 cells were serially re-stimulated with irradiated K562 targets three times, with absolute counts and viability assessments taken at regular intervals over 17 days. Cell counts and viability measurements were obtained using the LUNA Automated Cell Counter (Logos Biosystems). Supernatants were collected 24 h after each stimulation for longitudinal measurements of cytokine levels in cultures, as previously described<sup>41</sup>.

**Intracellular cytokine, perforin and granzyme B analysis.** CD8<sup>+</sup> T cells from Patient-10 were stimulated 3:1 with paramagnetic beads coated with anti-CD3 and anti-CD28 monoclonal antibodies for 6 h in the presence of CD107a monoclonal antibody and the golgi inhibitors brefeldin A and monensin. Cells were washed and stained with live/dead viability dye, followed by surface staining for CD3, CD8 and TCRVβ5.1. These lymphocytes were subsequently fixed/permeabilized and intracellularly stained for IFNγ. CAR T cells generated from healthy donors (*TET2* knockdown or control) were stimulated in the same way with CD3/CD28 beads or beads coated with an anti-idiotypic antibody against CAR19. Cells were then stained for surface antigens (CD3, CD4, CD8 and CAR19). After fixation and permeabilization, intracellular staining for IFNγ, TNFα and IL-2 was performed.

For perforin and granzyme B analysis, CTL019 cells that had been transduced with a *TET2* or scrambled control shRNA were expanded for 14 days and cryopreserved. These CAR T cells were then thawed and rested for 4 h, followed by live/dead and surface staining for CD3, CD8 and CAR19. Intracellular staining for perforin and granzyme was carried out following fixation and permeabilization. Cells were analysed on an LSRFortessa (BD).

**Cytotoxicity assay.** Healthy donor CTL019 cells transduced with a shRNA directed against *TET2* or a scrambled control were co-cultured with CBG luciferase-expressing NALM-6 and OSU-CLL cell lines at the indicated ratios for 16 h<sup>41</sup>. Cell extracts were created using the Bright-Glo Luciferase Assay System (Promega Corporation) and substrate was added according to the manufacturer's instructions. Luciferase measurements were taken on a SpectraMax luminescence microplate reader (Molecular Devices), and specific lysis was calculated as previously described<sup>41</sup>.

**Analysis of *TET2* gene expression levels in T cell subsets.** *TET2* expression levels were determined by analysing a published gene expression dataset (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23321>) for CD8<sup>+</sup> T cell subsets (naive, stem cell memory, central memory and effector memory) isolated from three healthy human subjects<sup>42</sup>. Genechip (Affymetrix) data were processed with the Bioconductor Oligo software package<sup>43</sup> (release 3.6, Bioconductor) using the RMA method<sup>44</sup>.

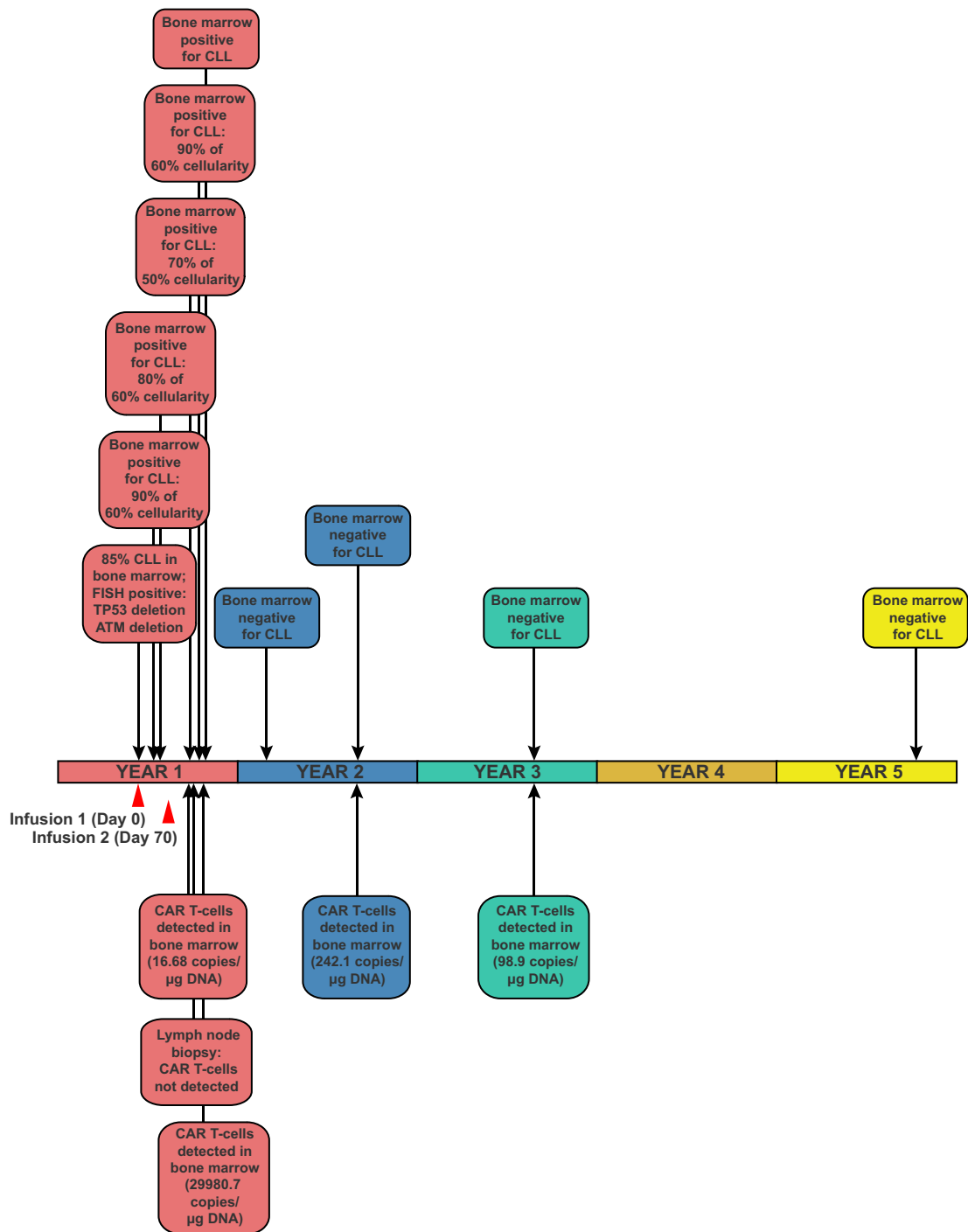
**Statistical analyses.** Normality was assessed for all data using the D'Agostino–Pearson omnibus test. For integration site data analysis, genomic feature data comparisons were carried out as previously described using χ<sup>2</sup>, Fisher's exact tests, or a combination of Bayesian model averaging, conditional logit and regression<sup>27,45,46</sup>. Assessments of T cell differentiation and function in shRNA-mediated *TET2* knockdown experiments were performed using a paired Student's *t*-test. Estimates of variation within each group of data are presented as error bars. Analyses were performed with SAS (SAS Institute), Stata 13.0 (StataCorp) or GraphPad Prism 6 (GraphPad Software). Cytokine heat maps were constructed using Morpheus software (Broad Institute; accessed at <https://software.broadinstitute.org/morpheus/>). All tests were two-sided. Exact *P* values are reported. *P* < 0.05 was considered statistically significant.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

**Data availability.** Sequencing data are available at the NCBI Sequence Read Archive (accession number SRP136348; analyses of lentiviral integration sites, *TET2* chimaeric transcripts and alleles as well as mutations associated with hematologic malignancies), Adaptive Biotechnologies' immuneACCESS database (<https://doi.org/10.21417/B72D1Q>; TCRVβ and BCR IgH immunosequencing) and Gene Expression Omnibus (accession number GSE112494; ATAC-seq). Any other data that support the findings of the study will be made available upon reasonable request.

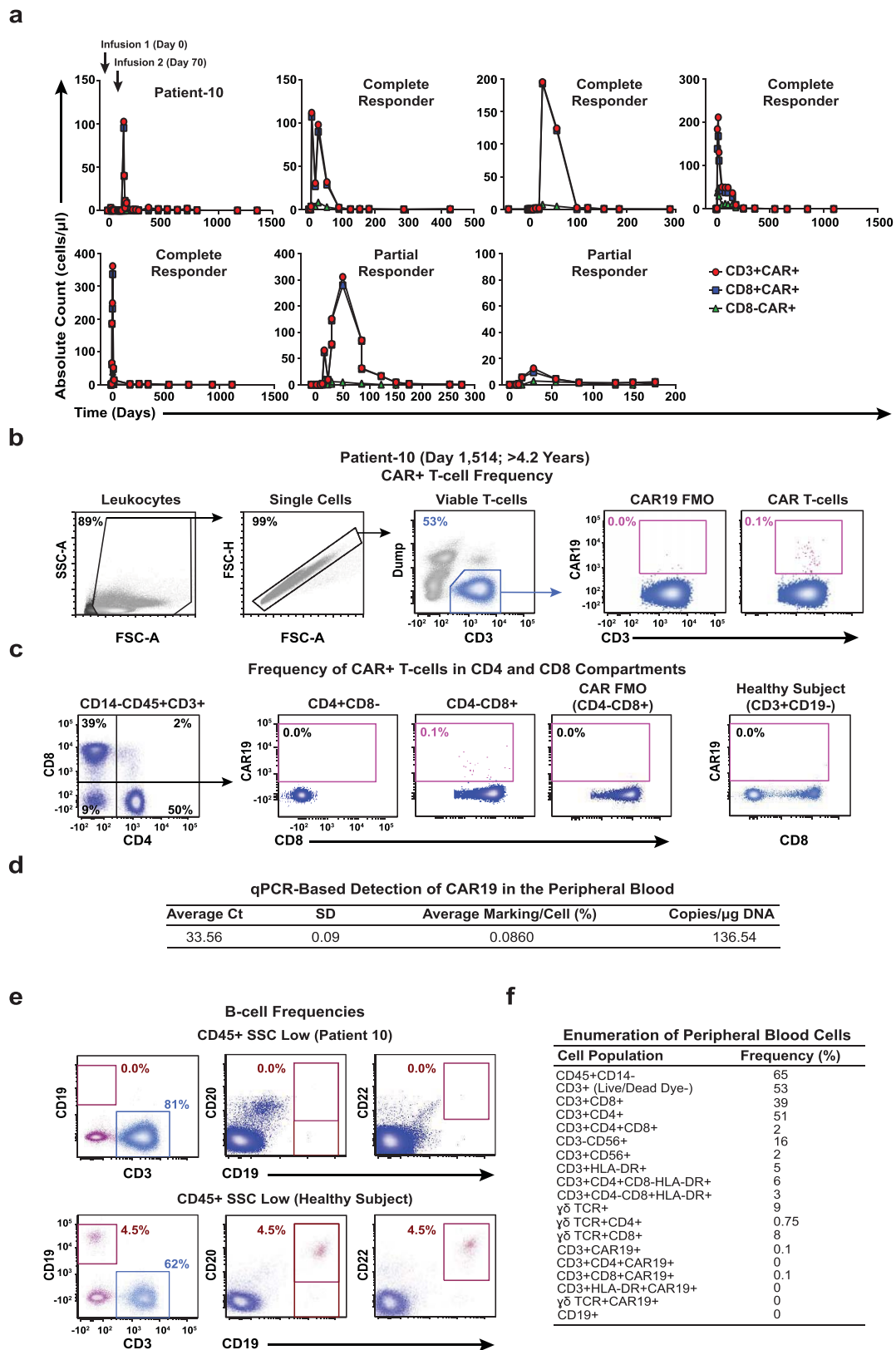


22. Hertlein, E. et al. Characterization of a new chronic lymphocytic leukemia cell line for mechanistic in vitro and in vivo studies relevant to disease. *PLoS ONE* **8**, e76607 (2013).
23. Barrett, D. M. et al. Treatment of advanced leukemia in mice with mRNA engineered T cells. *Hum. Gene Ther.* **22**, 1575–1586 (2011).
24. Maude, S. L. et al. Chimeric antigen receptor T cells for sustained remissions in leukemia. *N. Engl. J. Med.* **371**, 1507–1517 (2014).
25. Jena, B. et al. Chimeric antigen receptor (CAR)-specific monoclonal antibody to detect CD19-specific T cells in clinical trials. *PLoS ONE* **8**, e57838 (2013).
26. Brady, T. et al. A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.* **39**, e72 (2011).
27. Berry, C., Hannenhalli, S., Leipzig, J. & Bushman, F. D. Selection of target sites for mobile DNA integration in the human genome. *PLOS Comput. Biol.* **2**, e157 (2006).
28. Berry, C. C. et al. Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* **28**, 755–762 (2012).
29. Berry, C. C., Ocwieja, K. E., Malani, N. & Bushman, F. D. Comparing DNA integration site clusters with scan statistics. *Bioinformatics* **30**, 1493–1500 (2014).
30. Scholler, J. et al. Decade-long safety and function of retroviral-modified chimeric antigen receptor T cells. *Sci. Transl. Med.* **4**, 132ra53 (2012).
31. Daber, R., Sukhadia, S. & Morrisette, J. J. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. *Cancer Genet.* **206**, 441–448 (2013).
32. Liu, M. Y. et al. Mutations along a TET2 active site scaffold stall oxidation at 5-hydroxymethylcytosine. *Nat. Chem. Bio.* **13**, 181–187 (2017).
33. Hu, L. et al. Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545–1555 (2013).
34. Liu, M. Y., DeNizio, J. E. & Kohli, R. M. Quantification of oxidized 5-methylcytosine bases and TET enzyme activity. *Methods Enzymol.* **573**, 365–385 (2016).
35. Jin, J. et al. Simplified method of the growth of human tumor infiltrating lymphocytes in gas-permeable flasks to numbers needed for patient treatment. *J. Immunother.* **35**, 283–292 (2012).
36. Carty, S. A. et al. The loss of TET2 promotes CD8<sup>+</sup> T cell memory differentiation. *J. Immunol.* **200**, 82–91 (2018).
37. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
38. Pauken, K. E. et al. Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science* **354**, 1160–1165 (2016).
39. Xu, J. et al. Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells. *Nat. Genet.* **49**, 377–386 (2017).
40. Laport, G. G. et al. Adoptive transfer of costimulated T cells induces lymphocytosis in patients with relapsed/refractory non-Hodgkin lymphoma following CD34<sup>+</sup>-selected hematopoietic cell transplantation. *Blood* **102**, 2004–2013 (2003).
41. Fraietta, J. A. et al. Ibrutinib enhances chimeric antigen receptor T cell engraftment and efficacy in leukemia. *Blood* **127**, 1117–1127 (2016).
42. Gattinoni, L. et al. A human memory T cell subset with stem cell-like properties. *Nat. Med.* **17**, 1290–1297 (2011).
43. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
44. Irizarry, R. A. et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
45. Brady, T. et al. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009).
46. Ocwieja, K. E. et al. HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog.* **7**, e1001313 (2011).
47. Moskowitz, D. M. et al. Epigenomics of human CD8 T cell differentiation and aging. *Sci. Immunol.* **2**, eaag0192 (2017).



**Extended Data Fig. 1 | Timeline of disease clearance by CAR T cells in Patient-10.** An outline of clinical findings in Patient-10, including the

results of bone marrow assessments, tumour cytogenetics and CAR T cell persistence. CTL019 cell infusion time points are indicated by red arrows.



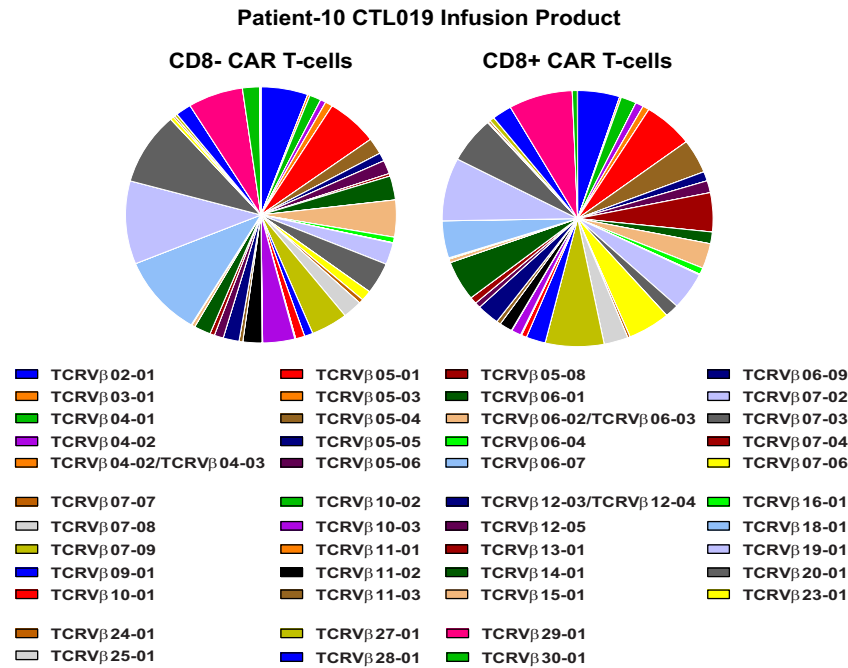
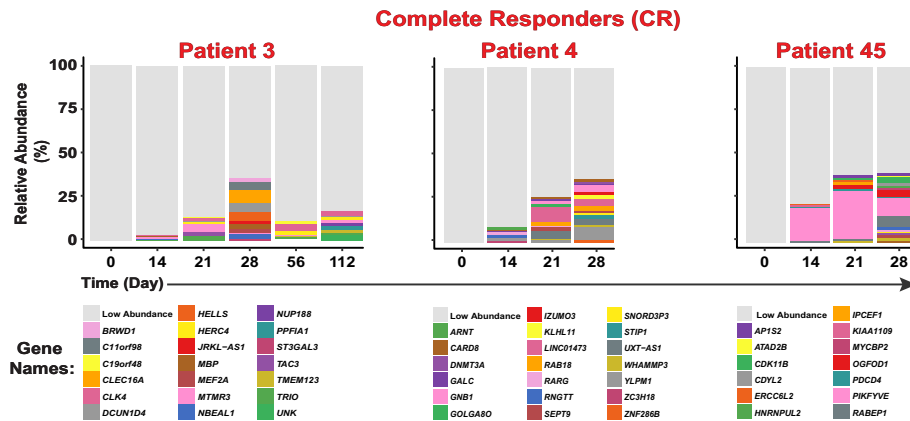
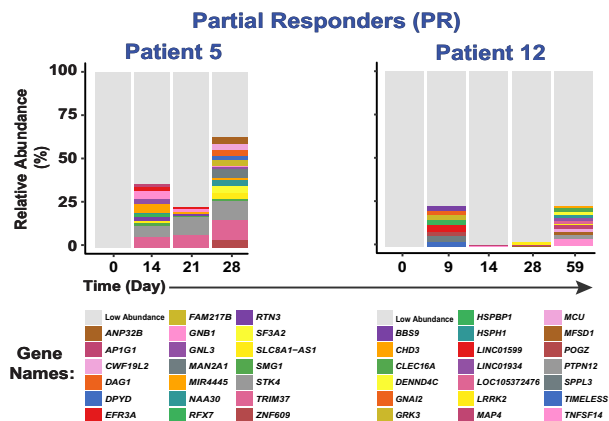
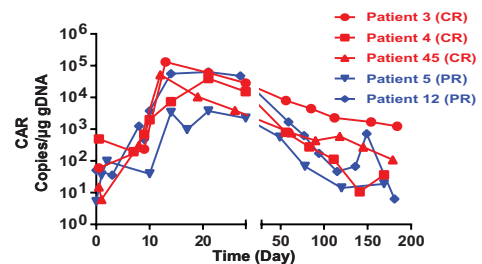
Extended Data Fig. 2 | See next page for caption.

**Extended Data Fig. 2 | CAR T cell detection and profiling of immune cell populations in Patient-10 and other responders.** **a**, Pre- and post-infusion kinetics of CAR T cell expansion ( $CD3^+$ ,  $CD8^+$  and  $CD8^-$ ) are shown in Patient-10 compared to other responders. The number of circulating CTL019 cells was calculated based on frequencies of  $CD3^+$ ,  $CD8^+$  and  $CD8^-$  CAR T cell populations and absolute cell counts.

**b**, Flow cytometry gating strategy to identify peripheral blood CAR T cells in Patient-10. **c**, Relative percentages of CTL019 cells in the CD4 and CD8 compartments of this patient. T cells from a healthy subject served as a negative control. **d**, The persistence of CAR T cells in the peripheral blood of Patient-10 was determined by qPCR. The average threshold cycle

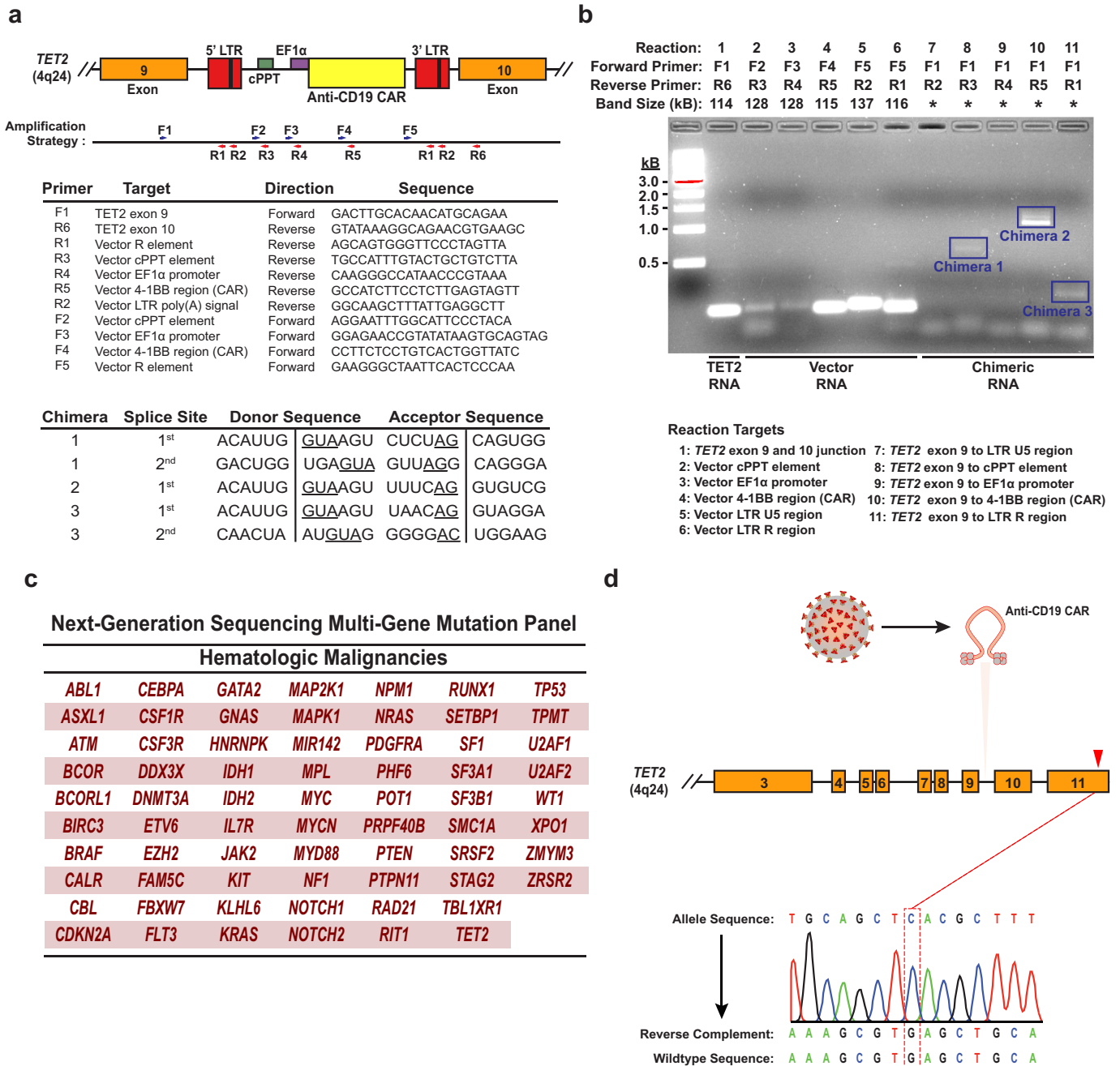
(Ct) value obtained from three replicates and a standard deviation (SD) is listed. Calculations of CAR T cell abundance are reported as average marking per cell as well as transgene copies per microgram of genomic DNA. **e**, Frequencies of circulating B cells in Patient-10 compared to a healthy subject. Pre-gating was performed to exclude dead cells as well as doublets, and all gating thresholds were based on fluorescence minus one (FMO) controls (representative of two independent experiments). **f**, Enumeration of various immune cell populations in the blood of Patient-10. The frequency of each population is listed in a separate column that corresponds to its phenotypic marker.



**a****b****c****d**

**Extended Data Fig. 3 | Clonal composition of CAR T cells from Patient-10 and other subjects treated with CTL019. a,** TCRVβ distribution in CD8<sup>-</sup> (left) and CD8<sup>+</sup> (right) CAR T cells in the cellular infusion product of Patient-10. **b,** Relative frequencies of CAR T cell clones in three patients who had complete responses to CTL019 therapy, summarized as stacked bar graphs. Each colour (horizontal bar) denotes a

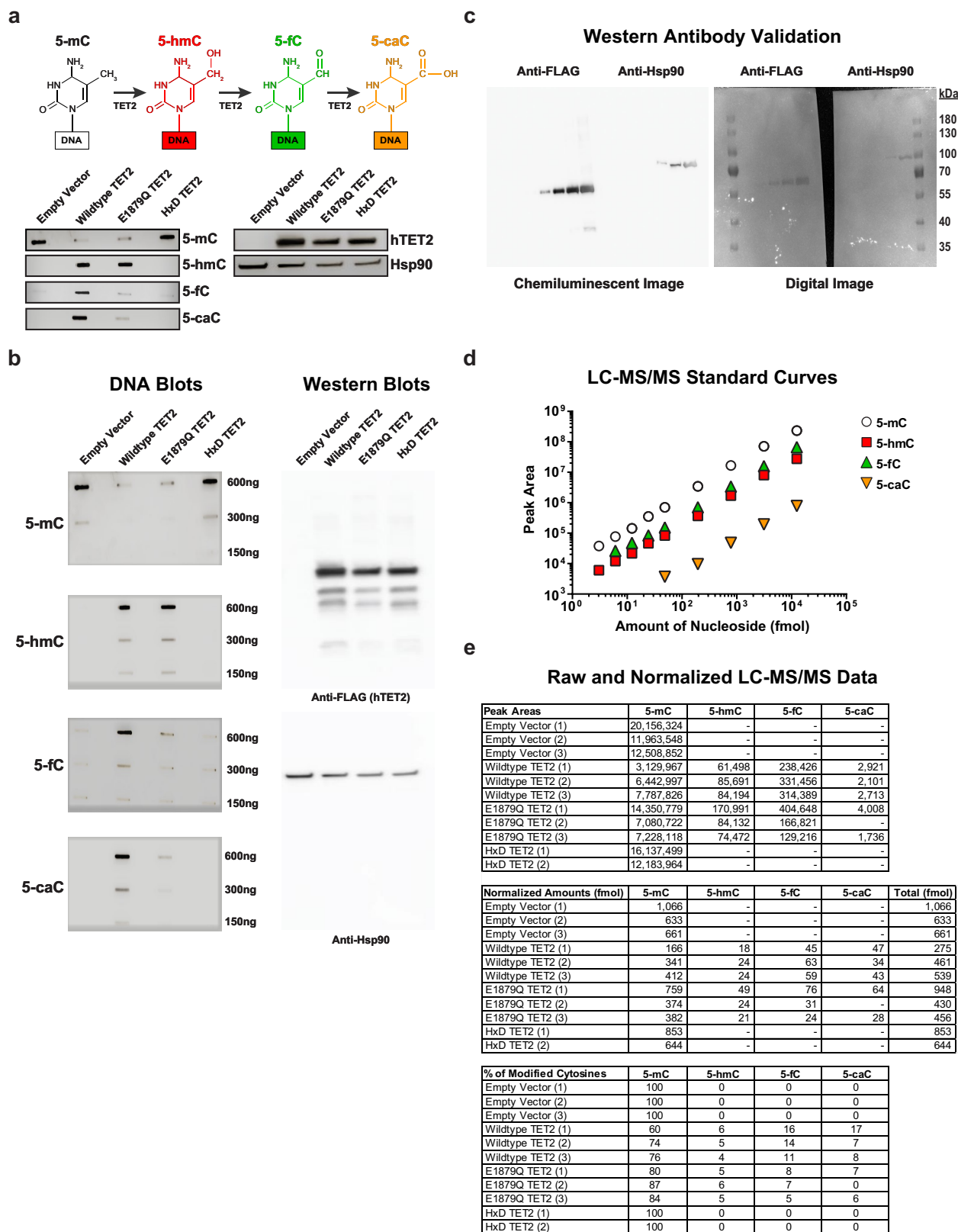
major cell clone, as marked by lentiviral integration sites. **c,** Integration site analysis in CAR T cells from two partially responding patients. **d,** In vivo expansion of CAR T cells in the above patients as determined by quantifying the average CAR transgene copies per microgram DNA at each time point.



#### Extended Data Fig. 4 | Detection of *TET2* chimaeric transcripts in Patient-10 CAR T cells and DNA sequencing for mutation detection.

**a**, The strategy for detection of polyadenylated RNA corresponding to truncated *TET2* transcripts is depicted. Boxes represent the genomic regions between *TET2* exons 9 and 10 with the integrated vector present. Blue and red arrows indicate general locations of the forward and reverse primers, which are listed below the diagram. LTR, long terminal repeat; cPPT, polypurine tract; EF1 $\alpha$ , elongation factor 1- $\alpha$  promoter. Sequences corresponding to the splice junctions for the three chimaeric messages (five total junctions) are listed in the bottom chart. Underlines indicate consensus splice donors and acceptors. **b**, Visualization of chimaeric *TET2* RT-PCR products. PCR products were separated on a native agarose

gel and stained with ethidium bromide. Expected sizes of amplicons are listed above the gel. Truncated transcripts are highlighted by blue boxes. A key to the RT-PCR reactions is shown below the diagram. \*Band size not determined (two independent experiments). **c**, Genes interrogated by the next generation sequencing panel used to analyse DNA isolated from CD8<sup>+</sup>CAR<sup>+</sup> T cells and CAR<sup>-</sup> T cells in Patient-10 at the peak of his response. **d**, Sanger sequencing of specific amplifications corresponding to the allele that was disrupted by integration of the CAR lentivirus is shown. The mutation that was detected by next generation sequencing of total genomic DNA from CAR<sup>+</sup> T cells (Fig. 3c) is not present in the *TET2* allele hosting the lentiviral integration site.

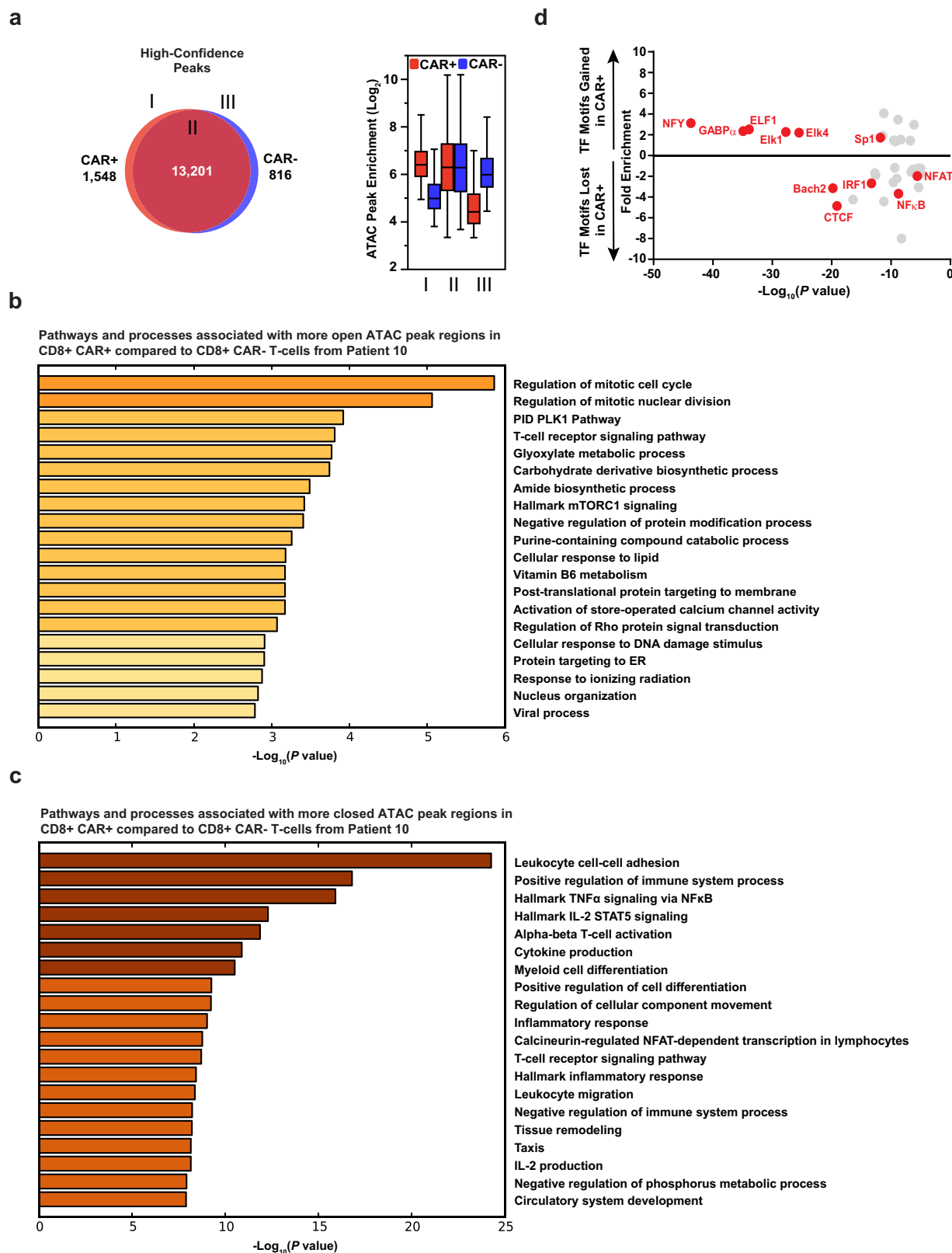


Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | Analysis of DNA methylation variants from HEK293T cells overexpressing TET2.** **a**, Depiction of sequential oxidations of 5-mC to 5-hmC, 5-fC and 5-caC catalysed by TET2 (top). Dot blots for 5-mC, 5-hmC, 5-fC and 5-caC in genomic DNA (gDNA) isolated from HEK293T cells transfected with the E1879Q TET2 mutant are shown. Assay controls include an empty vector, wild-type TET2 and a catalytically inactive (HxD) TET2 mutant (bottom left). A western blot using anti-FLAG antibody to detect hTET2 in the above cells is also shown. Hsp90 $\alpha/\beta$  was used as a loading control (bottom right). Brightness and contrast were adjusted evenly across blots. **b**, Original, uncropped DNA (left) and western (right) blots. A dilution series was used for semiquantitative analysis of DNA methylation variants. Representative results of three independent experiments are shown. **c**, Validation of anti-FLAG and anti-Hsp90 $\alpha/\beta$  antibodies. Serial dilutions of lysates (0.008, 0.04, 0.2, 0.8, 4.0  $\mu\text{g } \mu\text{l}^{-1}$ ) obtained from HEK293T cells transfected with wild-type TET2. Gels were probed by western

blot. Both chemiluminescence (left) and digital (right) images were captured, demonstrating that these antibodies exhibit specificity for the expected FLAG tag and Hsp90 based on molecular weight markers. **d**, 5-mC, 5-hmC, 5-fC and 5-caC nucleosides were analysed at fixed concentrations using LC-MS/MS to generate standard curves. The area under the curve (AUC) was calculated for each MS/MS fragment. In the case of 5-hmC, the slope was further adjusted because of quality control analysis of an equimolar mixture of oligonucleotides, each containing a single modification. **e**, Analysis of gDNA from individual biological replicates within each HEK293T cell group. The top chart lists the raw AUCs that were converted to relative amounts of modified cytosine (middle) according to their signal intensities from the respective standard curves. The percentage of each modified cytosine calculated for each sample is shown in the bottom chart. Results are from three independent experiments.



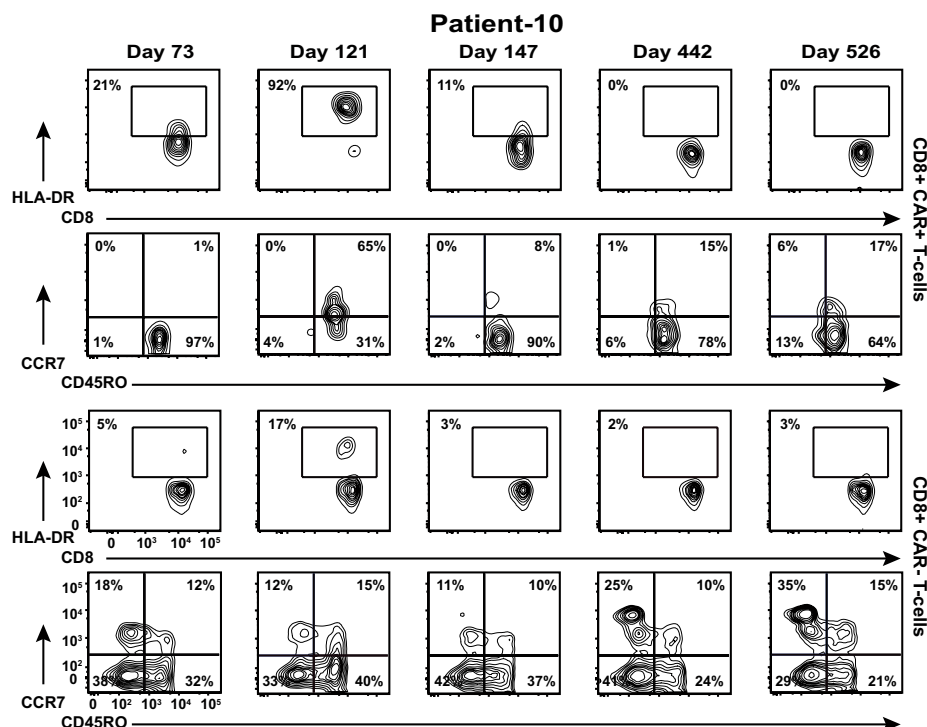


Extended Data Fig. 6 | See next page for caption.

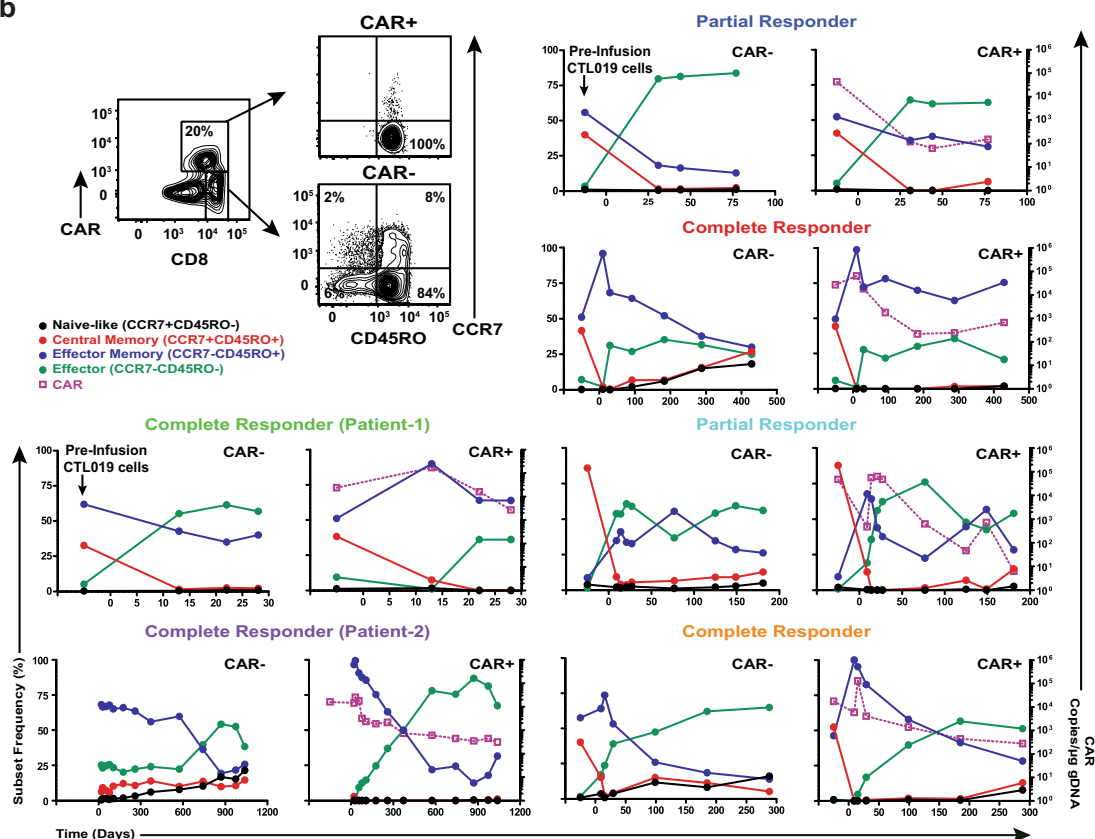
**Extended Data Fig. 6 | Global chromatin profiling of CAR<sup>+</sup> and CAR<sup>-</sup> T cells from Patient-10.** **a**, Venn diagrams of high-confidence differential ATAC-seq regions (left) and enrichment of those peaks in regions of the diagrams (right) in CAR<sup>+</sup> and CAR<sup>-</sup> CD8<sup>+</sup> T cells expanded from Patient-10 (two biological replicates analysed in two independent experiments). Boxes extend from the 25th to 75th percentiles, the middle line denotes the median and whiskers show minimum and maximum. **b**, Gene Ontology terms associated with chromatin regions that are significantly more open in CD8<sup>+</sup>CAR<sup>+</sup> T cells from Patient-10 compared to their matched CD8<sup>+</sup>CAR<sup>-</sup> T cell counterparts. **c**, Ontology analysis for chromatin regions that are less accessible in CD8<sup>+</sup>CAR<sup>+</sup> T cells than in CD8<sup>+</sup>CAR<sup>-</sup> T cells. **d**, Enrichment of transcription factor (TF) binding

motifs in chromatin regions gained or lost in CAR<sup>+</sup> compared to CAR<sup>-</sup> T cells from Patient-10. Transcription factor motifs that were potentially more accessible in increased ATAC-seq peaks of CAR<sup>+</sup> T cells included E26 transformation-specific (ETS) (GABP $\alpha$ , ELF1, Elk4) and zinc finger (ZF) transcription factor (Sp1) binding sites that are known to be enriched in human CD8<sup>+</sup> T cells before differentiation occurs<sup>47</sup>. Transcription factor motifs that were potentially less accessible owing to reduced ATAC-seq peaks in CAR<sup>+</sup> T cells from Patient 10 (NF- $\kappa$ B, IRF1, NFAT-AP1 and CTCF) are enriched in terminally differentiated effector and exhausted T cells and have known key roles in forming the epigenetic landscape that programs their biology<sup>38</sup>.

a

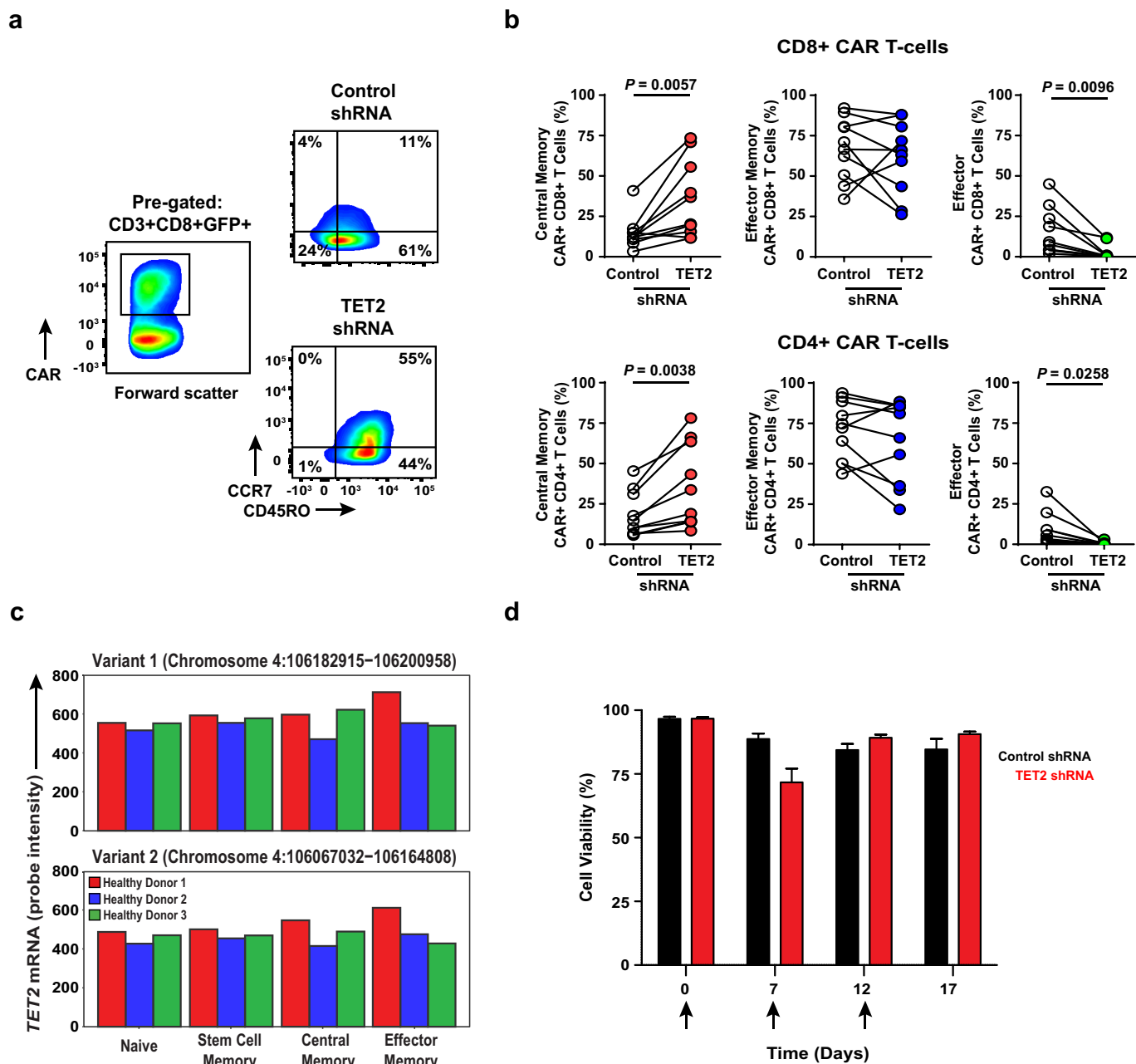


b



**Extended Data Fig. 7 | The differentiation state of CAR T cells in Patient-10 compared to other responders over time. a,** Representative contour plots of flow cytometric data depicting the frequency of CAR<sup>+</sup> and CAR<sup>-</sup>CD8<sup>+</sup> T cells in Patient-10 that express HLA-DR. The proportions of these cells that express CD45RO and CCR7 as determinants of differentiation status are shown. Contour plot insets indicate the

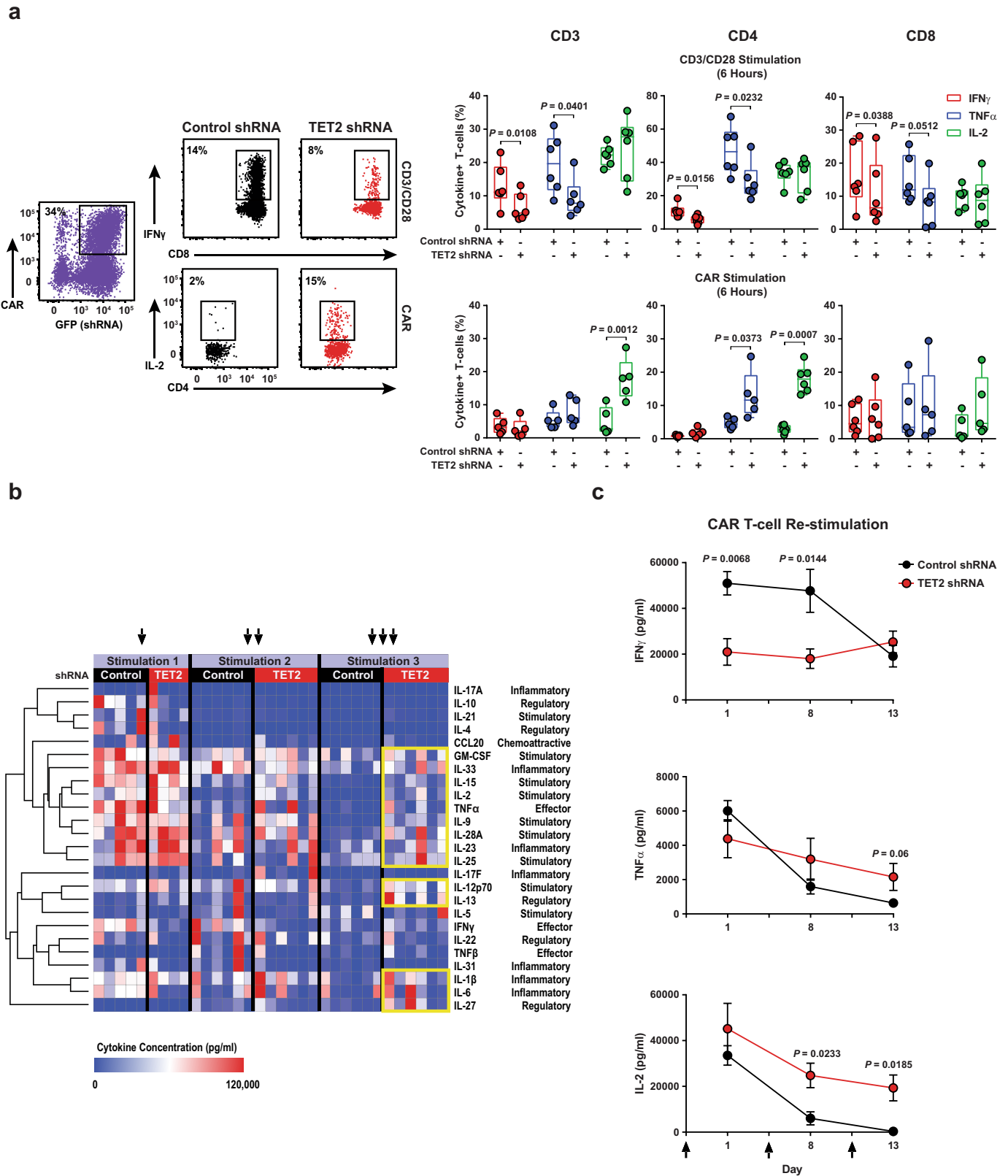
frequencies of the gated cell populations. **b,** Example gating strategy used to determine the differentiation phenotype of CD8<sup>+</sup>CAR<sup>+</sup> and CAR<sup>-</sup> T cells from a complete responder (top left). Line graphs depict the differentiation state of these cell populations in other responding patients over time and are plotted with corresponding CAR T cell levels in the blood, as determined by qPCR.



**Extended Data Fig. 8 | Effect of *TET2* expression on T cell differentiation and viability.** **a**, Representative flow cytometry plots showing the differentiation state of healthy donor  $CD8^+CAR^+$  T cells after transduction with a scrambled shRNA (control) or shRNA targeting *TET2*. Insets define frequencies of gated populations. **b**, Frequencies of healthy subject  $CAR^+CD8^+$  (top) and  $CAR^+CD4^+$  (bottom) T cells according to differentiation phenotype following control or *TET2* shRNA transduction ( $n = 10$ ; pooled results from four independent experiments). *P* values were determined using a two-tailed, paired Student's *t*-test. **c**, Comparison of

the expression levels of *TET2* in naive and memory  $CD8^+$  T cell subsets from three healthy donors. Two variants encoding different isoforms have been identified for this gene in humans. Expression levels of each *TET2* variant were estimated by measuring the probe intensity from microarray analysis. **d**, Viability of  $CAR^+$  T cells transduced with a *TET2* shRNA or scrambled control and restimulated with K562 cells expressing CD19 ( $n = 12$ ; pooled results from three independent experiments). Each arrow indicates the time point at which CAR T cells were exposed to antigen. Error bars depict s.e.m.



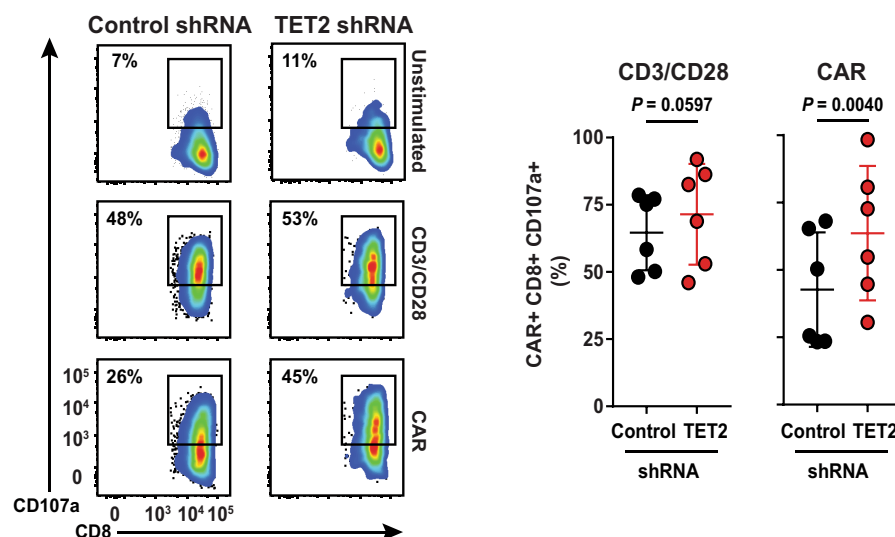


Extended Data Fig. 9 | See next page for caption.

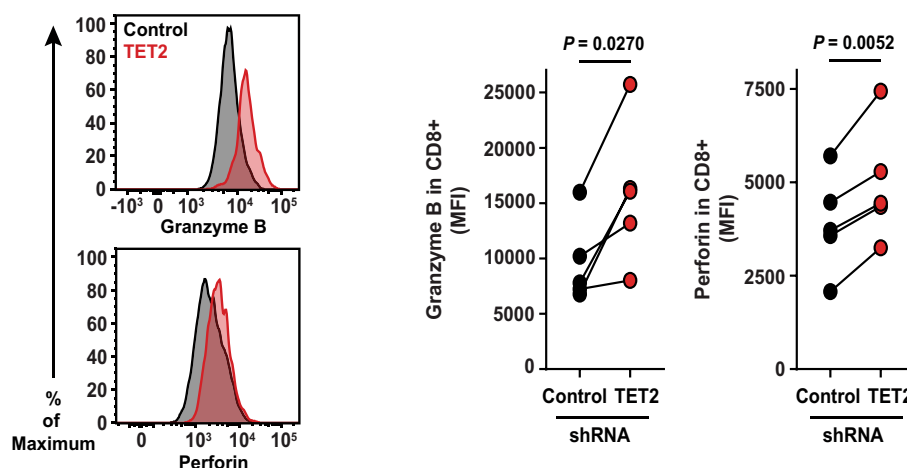
**Extended Data Fig. 9 | CAR T cell cytokine profiles following TET2 inhibition.** **a**, Representative flow cytometry of acute intracellular cytokine production by healthy donor ( $n = 6$ ; three independent experiments) CAR T cells transduced with a *TET2* shRNA or a scrambled control shRNA (left). Production of IFN $\gamma$ , TNF $\alpha$  and IL-2 by total CD3 $^{+}$ , CD4 $^{+}$  and CD8 $^{+}$  CAR T cells is shown. These cells were stimulated with beads coated with anti-CD3 and anti-CD28 antibodies (top right) or CAR anti-idiotypic antibodies (bottom right). Boxes represent the 25th to 75th percentiles, the middle line denotes the median and whiskers depict minimum and maximum. **b**, Heat map and cluster analysis of cytokine

profiles for CAR T cells transduced with a *TET2* shRNA or scrambled control and serially restimulated with irradiated K562 cells expressing CD19 are shown. Colours represent scaled cytokine data corresponding to each stimulation time point. Hierarchical clustering was used to generate the cluster dendrogram and cytokine response groups. **c**, Production of IFN $\gamma$  (top), TNF $\alpha$  (middle) and IL-2 (bottom) by *TET2* knockdown or control CAR T cells ( $n = 6$ ; three independent experiments) following restimulation with CD19 antigen. Black arrows indicate when CAR T cells were exposed to CD19-expressing K562 cells. Error bars denote s.e.m. All *P* values were determined using a two-tailed, paired Student's *t*-test.

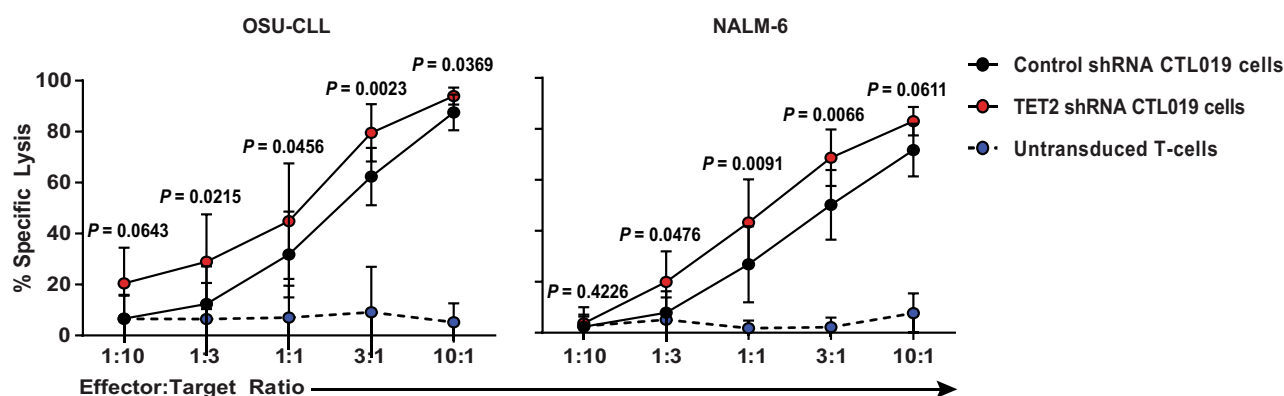
a



b

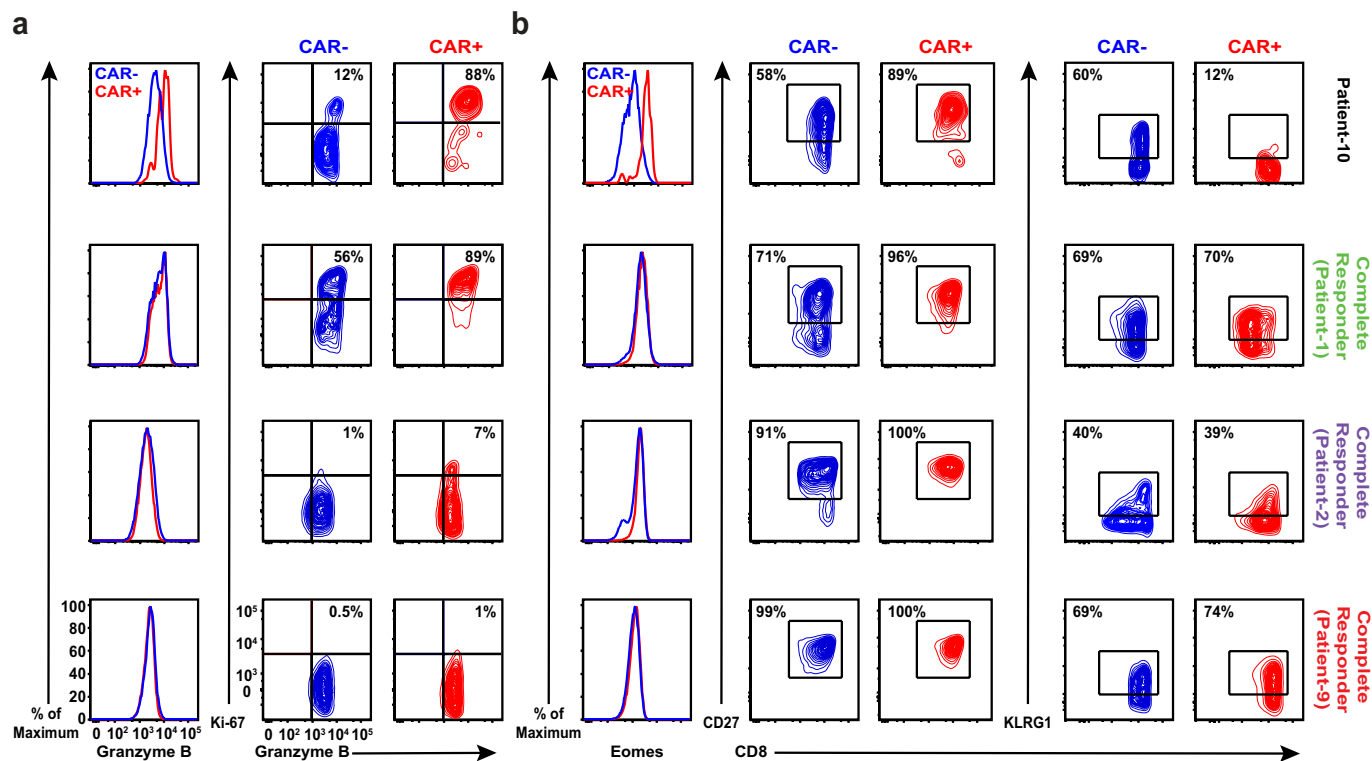


c



**Extended Data Fig. 10 | Effect of TET2 knockdown on the cytotoxic machinery of CAR T cells.** a, Flow cytometry plots showing the frequency of TET2 knockdown or control CAR T cells expressing CD107a (a marker of cytotoxicity) following CD3 and CD28 or CAR-specific stimulation (left). Summarized data from analysis of CAR T cells manufactured from  $n = 6$  different healthy donors is shown (right). b, Representative histograms illustrating expression levels of granzyme B and perforin in CAR T cells in the setting of TET2 inhibition as compared to its counterpart control (left).

Pooled data from CAR T cells of  $n = 5$  healthy donors are summarized on the right. c, Cytotoxic capacity of CTL019 cells (transduced with a TET2 or scrambled control shRNA) after overnight co-culture with luciferase-expressing OSU-CLL (left) or NALM-6 (right) cells. Untransduced T cells were included as an additional group to control for non-specific lysis. P-values were determined using a two-tailed, paired Student's *t*-test. All data were pooled from three independent experiments.



**Extended Data Fig. 11 | Effector and memory molecule expression by CAR T cells from Patient-10 compared to those from other responding subjects. a,** Expression of granzyme B (left) and the frequency of CAR<sup>-</sup> and CAR<sup>+</sup> T cells co-expressing granzyme B/Ki-67 (right panel) at the peak of in vivo CTL019 expansion in Patient-10 compared to three other

complete responders. **b,** Representative histograms of intracellular Eomes expression (left), and contour plots depicting frequencies of CD27<sup>-</sup> (middle) and KLRG1-expressing (right) lymphocytes in the same cell populations of these patients. These results are representative of three experiments repeated independently with comparable findings.



# Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II

Huasong Lu<sup>1,2</sup>, Dan Yu<sup>2</sup>, Anders S. Hansen<sup>2</sup>, Sourav Ganguly<sup>2</sup>, Rongdiao Liu<sup>1</sup>, Alec Heckert<sup>2</sup>, Xavier Darzacq<sup>2</sup> & Qiang Zhou<sup>2\*</sup>

**Hyperphosphorylation of the C-terminal domain (CTD) of the RPB1 subunit of human RNA polymerase (Pol) II is essential for transcriptional elongation and mRNA processing<sup>1–3</sup>. The CTD contains 52 heptapeptide repeats of the consensus sequence YSPTSPS. The highly repetitive nature and abundant possible phosphorylation sites of the CTD exert special constraints on the kinases that catalyse its hyperphosphorylation. Positive transcription elongation factor b (P-TEFb)—which consists of CDK9 and cyclin T1—is known to hyperphosphorylate the CTD and negative elongation factors to stimulate Pol II elongation<sup>1,4,5</sup>. The sequence determinant on P-TEFb that facilitates this action is currently unknown. Here we identify a histidine-rich domain in cyclin T1 that promotes the hyperphosphorylation of the CTD and stimulation of transcription by CDK9. The histidine-rich domain markedly enhances the binding of P-TEFb to the CTD and functional engagement with target genes in cells. In addition to cyclin T1, at least one other kinase—DYRK1A<sup>6</sup>—also uses a histidine-rich domain to target and hyperphosphorylate the CTD. As a low-complexity domain, the histidine-rich domain also promotes the formation of phase-separated liquid droplets in vitro, and the localization of P-TEFb to nuclear speckles that display dynamic liquid properties and are sensitive to the disruption of weak hydrophobic interactions. The CTD—which in isolation does not phase separate, despite being a low-complexity domain—is trapped within the cyclin T1 droplets, and this process is enhanced upon pre-phosphorylation by CDK7 of transcription initiation factor TFIIB<sup>1–3</sup>. By using multivalent interactions to create a phase-separated functional compartment, the histidine-rich domain in kinases targets the CTD into this environment to ensure hyperphosphorylation and efficient elongation of Pol II.**

Among all transcription-related cyclins, cyclin (CYC) T—which includes T1 (CYCT1) and T2—has the longest C-terminal regions. In CYCT1, the N-terminal region is structured and contains the cyclin-box repeats required for binding and activating CDK9, whereas the C-terminal region has only a few isolated motifs and is mostly unstructured (Fig. 1a).

To determine whether the CYCT1 C-terminal region is important for regulating CDK9 activity, we performed in vitro kinase reactions to examine affinity-purified CDK9–CYCT1–Flag heterodimers that contained various truncated forms of CYCT1 with progressively shortened C termini to phosphorylate the CTD (Fig. 1a and Extended Data Fig. 1a). A mixture of glutathione S-transferase (GST)–CTD<sub>52</sub> (RPB1 CTD containing all 52 repeats) and GST–CTD<sub>9</sub> (CTD containing 9 consensus repeats) was used in all reactions as the kinase substrates.

The various CYCT1 truncations did not substantially affect the ability of the associated CDK9 to autophosphorylate by producing the ATP-dependent mobility shift (Fig. 1a), nor did the truncations decrease the phosphorylation of CTD<sub>9</sub>. However, upon truncation to a position at and beyond 533, CDK9 became largely unable to produce the hyperphosphorylated CTD<sub>52</sub> (hereafter, Ilo) as revealed by the anti-phospho-Ser5 antibody 3E8 (Fig. 1a). A similar pattern was also detected with

the anti-phospho-Ser2 antibody 3E10 (Extended Data Fig. 1b). Thus, a region around position 533 in CYCT1 promoted the hyperphosphorylation by CDK9 of CTD<sub>52</sub>.

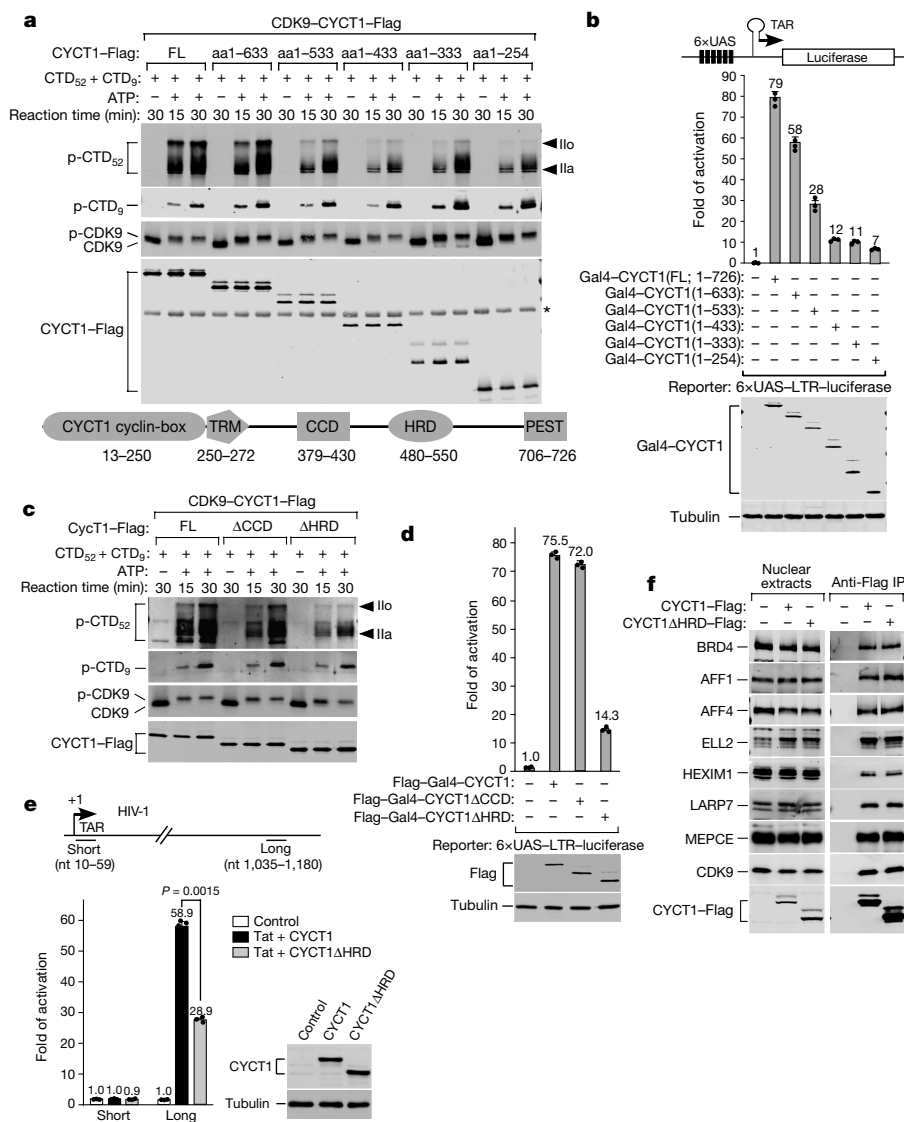
To determine whether the CYCT1 C-terminal truncations affect the transcriptional activity of P-TEFb, we used the Gal4-tethering system to test fusion proteins containing the Gal4 DNA-binding domain attached to the various truncated forms of CYCT1 to activate luciferase expression from the HIV-1 promoter containing Gal4-binding sites (Fig. 1b). Correlating with the kinase results, progressive CYCT1 C-terminal truncations to and beyond position 533 markedly reduced P-TEFb transcriptional activity (Fig. 1b).

The coiled-coil domain (CCD, amino acids 379–430) and the histidine-rich domain (HRD, amino acids 480–550) exist around position 533 (Fig. 1a). In kinase reactions, only the deletion of HRD ( $\Delta$ HRD) but not CCD ( $\Delta$ CCD) blocked the hyperphosphorylation by CDK9 of CTD<sub>52</sub> (Fig. 1c). Consistently, HRD but not CCD was required for P-TEFb transcriptional activity in the Gal4-tethering assay (Fig. 1d). Dependence on HRD for Tat/P-TEFb-activated HIV-1 elongation has previously been proposed<sup>7</sup>, and here confirmed by our finding that wild-type CYCT1 but not CYCT1 $\Delta$ HRD could effectively rescue the RNAi-knockdown of endogenous CYCT1 expression to produce long (promoter-distal) but not short (promoter-proximal) viral transcripts (Fig. 1e).

The importance of the HRD for the transcriptional activity of P-TEFb was generalized to cellular genes, as CYCT1 $\Delta$ HRD produced significantly less mRNA from four representative immediate early genes—*FOS*, *JUNB*, *MYC* and *EGR1*—as well as *HSP70-1* (also known as *HSPA1A*) under both basal and heat-shock conditions (Extended Data Fig. 2), all of which require P-TEFb for optimal transcription. Co-immunoprecipitation analysis (Fig. 1f and Extended Data Fig. 2e) shows that the decreased activity of CYCT1 $\Delta$ HRD was not due to any substantial change in binding to the major P-TEFb partners, including BRD4 and subunits of the super elongation complex and the 7SK small nuclear ribonucleoprotein particle that regulate P-TEFb activity<sup>1,4,5</sup>.

What is the mechanism by which the HRD promotes the activity of P-TEFb? The first hint came from the finding that wild-type CYCT1 in HeLa nuclei was more resistant to salt extraction, as compared to CYCT1 $\Delta$ HRD (Fig. 2a), which suggests that the HRD promoted the retention of CYCT1 in the nucleus. To determine more precisely the location and dynamics of this retention, we performed fluorescence recovery after photobleaching (FRAP) analysis of bindings of Halo-tagged wild-type CYCT1 and CYCT1 $\Delta$ HRD to a gene array activated by reverse tetracycline-controlled transactivator (Fig. 2b). This array contains about 200 copies of an integrated transgene marked by YFP–Lac repressor bound to the *lac* operator, which enables FRAP to be performed at this spot<sup>8</sup>. We observed significantly faster FRAP recovery for CYCT1 $\Delta$ HRD on the array than for wild-type CYCT1 (Fig. 2c). Further quantitative analysis (Extended Data Fig. 3a, b and Supplementary Information) revealed that this was mainly due to longer apparent residence time of wild-type CYCT1 ( $\tau_{\text{off}} \sim 56$  s), as compared to CYCT1 $\Delta$ HRD ( $\tau_{\text{off}} \sim 4$  s).

<sup>1</sup>School of Pharmaceutical Sciences, Xiamen University, Xiamen, China. <sup>2</sup>Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA. \*e-mail: qzhou@berkeley.edu



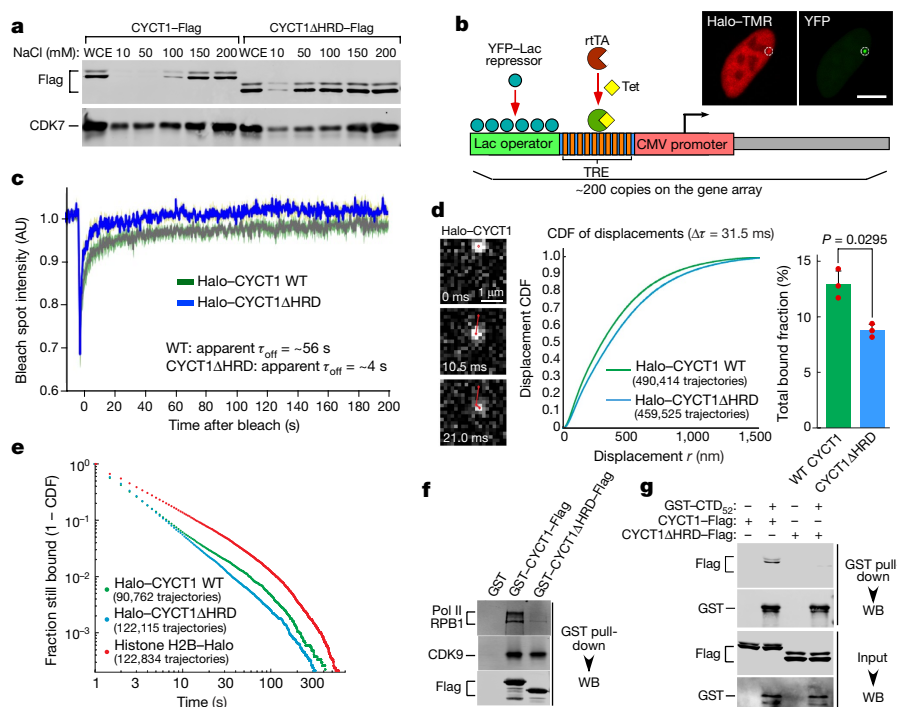
**Fig. 1 | The CYCT1 HRD promotes hyperphosphorylation of Pol II CTD<sub>52</sub> and activation of transcription, but not interaction with P-TEFb partners, by CDK9. **a**, **c**, P-TEFb containing the indicated CYCT1-Flag proteins (domain structure at bottom) were tested in kinase reactions containing GST-CTD<sub>52</sub> and GST-CTD<sub>9</sub> as the substrates. Western blotting was performed to detect the phosphorylated products (p-CTD<sub>52</sub> and p-CTD<sub>9</sub>) and P-TEFb components. Ilo and Ila, hyper- and hypo-phosphorylated CTD<sub>52</sub>, respectively. \*, a non-specific band. **b**, **d**, Plasmids expressing the indicated Gal4-CYCT1 fusions were co-transfected into HeLa cells with a HIV-1 LTR-luciferase reporter construct containing the Gal4 upstream activation sequences (UAS). Luciferase activities (mean  $\pm$  s.d.,  $n = 3$ ) in cell extracts were measured and compared to the activity in the first lane (set to 1). Levels of indicated proteins in extracts**

were examined by western blotting. **e**, HeLa-based CYCT1 knockdown cells containing an integrated HIV-1 provirus were transfected with the indicated Tat + CYCT1-expressing plasmids or a control vector. The indicated proteins were detected by western blotting. HIV-1 mRNA levels at the promoter-proximal 'short' and promoter-distal 'long' positions were analysed by qRT-PCR (mean  $\pm$  s.d.,  $n = 3$ ), with signals normalized to those in control cells (set to 1) and  $P$  value from two-tailed Student's  $t$ -test. **f**, Nuclear extracts of HeLa cells expressing the indicated CYCT1-Flag proteins and anti-Flag immunoprecipitates from nuclear extracts were analysed by western blotting. All western blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1. aa, amino acids; FL, full length; TRM, Tat/TAR-recognition motif; PEST, proline, glutamic acid, serine and threonine-rich sequence; nt, nucleotides.

To generalize the FRAP finding to endogenous genes (active and inactive) and cross-validate it using an orthogonal technique, we performed single-particle tracking in U2OS cells of Halo-tagged wild-type CYCT1 and CYCT1 $\Delta$ HRD proteins at 95 Hz to determine whether the HRD affects CYCT1 diffusion dynamics and bound fractions<sup>9,10</sup> (Fig. 2d, Extended Data Fig. 3c–e, Supplementary Video 1 and Supplementary Information). Consistent with the HRD-facilitated association of CYCT1 with chromatin and the transcriptional machinery, CYCT1 $\Delta$ HRD diffused faster and showed a smaller bound fraction (8.8%) than did wild-type CYCT1 (13.0%). Next, we conducted single-particle tracking using a longer exposure time (500 ms; Supplementary Video 2) to blur out fast-diffusing molecules and focus on bound molecules that were presumably located at predominantly

activated genes. We observed a wide distribution of CYCT1 binding events that probably encompassed both specific and non-specific interactions (Fig. 2e). Because this distribution did not fit well with models that assumed one or two single rate-limiting steps, we could not reliably attribute a single residence time. Nevertheless, consistent with the FRAP result, wild-type CYCT1 showed binding events that were significantly more stable than those of CYCT1 $\Delta$ HRD (Fig. 2e). Together, the single-particle tracking and FRAP results underscore the key role of the HRD in promoting the binding of P-TEFb to activated genes to phosphorylate CTD<sub>52</sub>.

What could be the direct target of the HRD on activated genes? Consistent with a previous report that a CYCT1 mutant that lacked the HRD failed to bind the CTD<sup>7</sup>, wild-type GST-CYCT1



**Fig. 2 | CYCT1 HRD contributes to the stable engagement of P-TEFb with activated gene array and endogenous genes, as well as direct binding to CTD by P-TEFb.** **a**, Nuclei of HeLa cells expressing indicated CYCT1-Flag were extracted with increasing NaCl concentrations. The whole cell extract (WCE) and soluble fractions were examined by western blotting. **b**, Diagram of the gene array analysed by FRAP in **c**. Tet, tetracycline; rtTA, reverse tetracycline-controlled transactivator; TRE, tetracycline response element. The fluorescence images show the Halo-TMR and YFP emissions collected in separate channels. The small circle indicates the bleach spot. **c**, Fluorescence recovery plots of indicated Halo-CYCT1 photobleached at the spot containing the activated gene array. The intensities were normalized to pre-bleach values and are shown at various time points after the bleach. Inferred residence times are also given. AU, arbitrary unit. **d**, Single-particle tracking (1-ms excitation pulse; 95 Hz) of Halo-tagged wild-type CYCT1 (21 cells) or CYCT1 $\Delta$ HRD (21 cells) labelled with photoactivatable Janelia Fluor 549 in U2OS cells. Left,

examples of raw images with trajectory overlaid in red. Middle, cumulative distribution function (CDF) of displacements at  $\Delta t = 31.5$  ms. Right, bound fraction inferred from three-state model fitting (mean  $\pm$  s.e.m.,  $n = 3$ .  $P$  value from two-tailed Student's  $t$ -test). **e**, Distribution of single-molecule binding times, uncorrected for photobleaching. Bound molecules (Janelia Fluor 646) were tracked using a long 500-ms exposure time. Plot shows 1 - CDF for trajectories captured for at least two frames. The constitutively bound histone H2B-Halo shows the photobleaching limit. Data from three independent replicates were merged and plotted. **f**, Immobilized GST or GST fusions were incubated with HeLa nuclear extract. The bound proteins were detected by western blotting (WB). **g**, Immobilized GST-CTD<sub>52</sub> was incubated with the indicated affinity-purified CYCT1-Flag. The bound proteins and 2.5% of input were analysed by western blotting. All western blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.

precipitated more Pol II from HeLa nuclear extracts than did GST-CYCT1 $\Delta$ HRD (Fig. 2f). Moreover, GST-CTD pulled down P-TEFb containing wild-type CYCT1, but not CYCT1 $\Delta$ HRD (Fig. 2g). Finally, a direct and HRD-dependent interaction was detected between a recombinant CYCT1 C-terminal fragment and mCherry-CTD in both the GST pulldown and glycerol gradient formats (Extended Data Fig. 8a, b). Thus, P-TEFb uses the HRD to directly target the CTD.

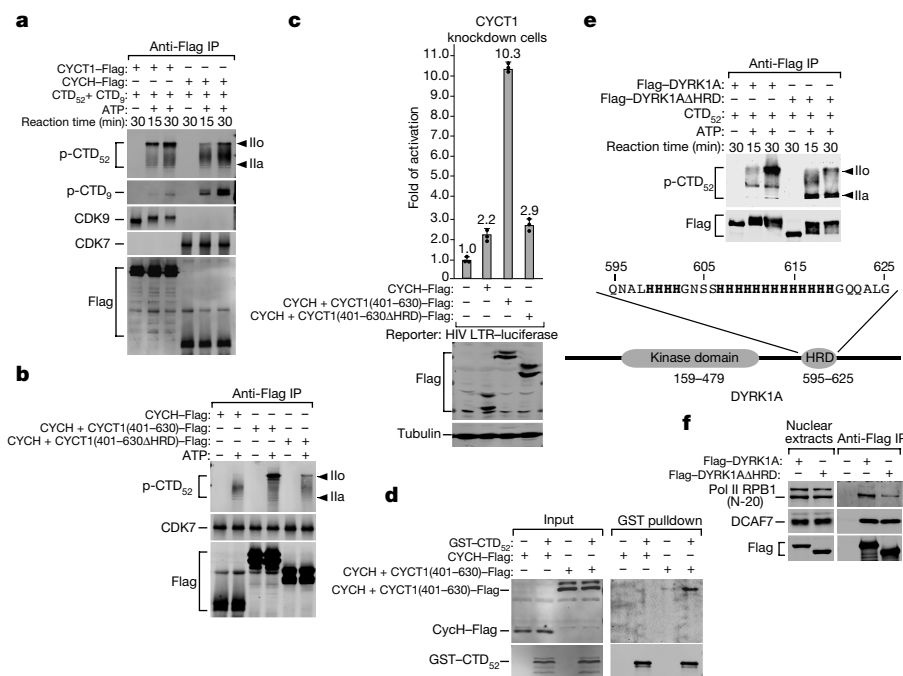
In addition to P-TEFb, CDK7-CYCH in TFIIF is also a CTD kinase for Pol II to clear the promoter during transcription initiation<sup>1</sup>. Unlike CYCT1, CYCH lacks a long C-terminal region and HRD. Compared to CDK9-CYCT1-Flag, affinity-purified CDK7-CYCH-Flag produced a markedly lower level of CTD<sub>52</sub> Ilo in kinase reactions (Fig. 3a), which was confirmed with recombinant CAK (CDK7-CYCH-MAT1) and P-TEFb (Millipore) in a time- and dosage-dependent manner (Extended Data Fig. 4a, b).

Appending a CYCT1 C-terminal fragment (amino acids 401–630) containing the HRD to CYCH markedly increased hyperphosphorylation by CDK7 of CTD<sub>52</sub> (Fig. 3b), and partially rescued the CYCT1 knockdown to support HIV-1 transcription (Fig. 3c). However, when the HRD was deleted from the CYCH + CYCT1 chimaera, both the production of CTD<sub>52</sub> Ilo and the rescue of CYCT1 knockdown were mostly abolished (Fig. 3b, c), which indicates the importance of the HRD to these processes. Consistently, the chimaera, but not wild-type CYCH, was precipitated by GST-CTD (Fig. 3d).

We next investigated whether other CTD kinases also use an HRD to target Pol II for hyperphosphorylation. We noticed that the kinase DYRK1A, a CTD kinase that controls transcription of selected growth-related genes<sup>6</sup> and that is associated with Down syndrome, also contains an HRD (Fig. 3e). The rest of DYRK1A and CYCT1 are non-homologous. In kinase reactions, wild-type DYRK1A and DYRK1A $\Delta$ HRD autophosphorylated to a similar extent, but only wild-type DYRK1A efficiently hyperphosphorylated CTD<sub>52</sub> (Fig. 3e). In a co-immunoprecipitation assay employing three different anti-RPB1 antibodies, wild-type DYRK1A precipitated more Pol II than did DYRK1A $\Delta$ HRD (Fig. 3f and Extended Data Fig. 4c). The precipitation of DCAF7, which binds DYRK1A N terminus<sup>11</sup>, was unaffected by DYRK1A $\Delta$ HRD. Thus, in addition to CYCT1, at least one other kinase also requires a functional HRD to efficiently bind and hyperphosphorylate CTD<sub>52</sub>.

The HRD is a low-complexity domain owing to the overrepresentation of only a single amino acid. This domain, including a central cluster of multiple consecutive histidines, is highly conserved in vertebrate CYCT1 (Extended Data Fig. 5a). Using the prediction program IUPred<sup>12</sup> (<http://iupred.enzim.hu>), we found that the human CYCT1 HRD displays the highest disorder tendency in a broader intrinsically disordered region (IDR) that lacks well-defined structure<sup>13</sup> and overlaps with the CYCT1 C-terminal region (Extended Data Fig. 5b). Recently, the IDRs—especially those containing a low-complexity domain—have been shown to promote liquid–liquid phase separation,





**Fig. 3 | Upon fusion with CYCT1 HRD, CYCH promotes CDK7 hyperphosphorylation of Pol II CTD; DYRK1A also contains a functional HRD to target and hyperphosphorylate CTD.**

**a, b, e,** The indicated anti-Flag immunoprecipitates (IP) were tested in kinase reactions as in Fig. 1a. **c,** The CYCT1 knockdown cells were co-transfected with HIV-1 LTR-luciferase reporter construct and plasmids expressing the indicated proteins. Luciferase activities were measured and

analysed as in Fig. 1b. Data are mean  $\pm$  s.d.,  $n = 3$ . **d,** Immobilized GST-CTD<sub>52</sub> was incubated with nuclear extract of HeLa cells expressing the indicated proteins. The bound proteins and 2.5% of input were analysed by western blotting. **f,** Nuclear extract of HeLa cells expressing the indicated proteins and anti-Flag immunoprecipitates from nuclear extract were analysed by western blotting. All western blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.

which probably drives the formation of intracellular membrane-less organelles<sup>14</sup> for compartmentalized biochemical reactions<sup>13</sup>.

In vitro, phase separation is reversible<sup>13</sup> and influenced by several parameters (for example, temperature, ionic strength, post-translational modifications and so on), which enable liquid droplets to form upon reaching a certain threshold and to quickly disassemble when pushed in the opposite direction<sup>15</sup>. To determine whether T1-IDR (amino acids 462–654)—the longest, HRD-containing CYCT1 IDR—could phase separate in an HRD-dependent manner, we purified the GFP-T1-IDR and GFP-T1-IDRΔHRD fusions from *Escherichia coli* (Extended Data Fig. 5c). At 150 mM NaCl, the protein solutions remained translucent. When lowered to 37.5 mM, a typical concentration used to induce phase separation in vitro<sup>16,17</sup>, the wild-type GFP-T1-IDR solution immediately turned opaque, whereas GFP-T1-IDRΔHRD showed no change (Extended Data Fig. 5e).

Under a microscope, wild-type GFP-T1-IDR spontaneously formed micrometre-sized, spherical droplets, whereas GFP-T1-IDRΔHRD produced only a low level of irregular aggregates (Fig. 4a). Similarly, the GFP fusion that contained the longest IDR (amino acids 491–686) of DYRK1A also formed droplets in an HRD-dependent manner (Extended Data Fig. 6). Notably, the GFP-T1-IDR droplets quickly disappeared once NaCl returned to 150 mM (Extended Data Fig. 5f). Furthermore, 1,6-hexanediol, a compound that is known to perturb weak hydrophobic interactions to disassemble structures that exhibit liquid-like properties<sup>18,19</sup>, completely blocked droplet formation (Extended Data Fig. 5g). Finally, when nine histidines within the histidine cluster in CYCT1 HRD were changed to alanines (Extended Data Fig. 7a), the resulting GFP-T1-IDR(9A) mutant formed only tiny droplets (Fig. 4b), which indicates that the histidines are essential for phase separation. Functionally, the mutation decreased phosphorylation by P-TEFb of CTD<sub>52</sub>, and the transcriptional activity of P-TEFb (Extended Data Fig. 7c, d).

In cells, both ectopically expressed CYCT1-Flag or eGFP-CYCT1 and endogenous CYCT1 displayed a punctuated staining pattern inside the nuclei (Fig. 4c–e and Extended Data Fig. 5h). This has previously

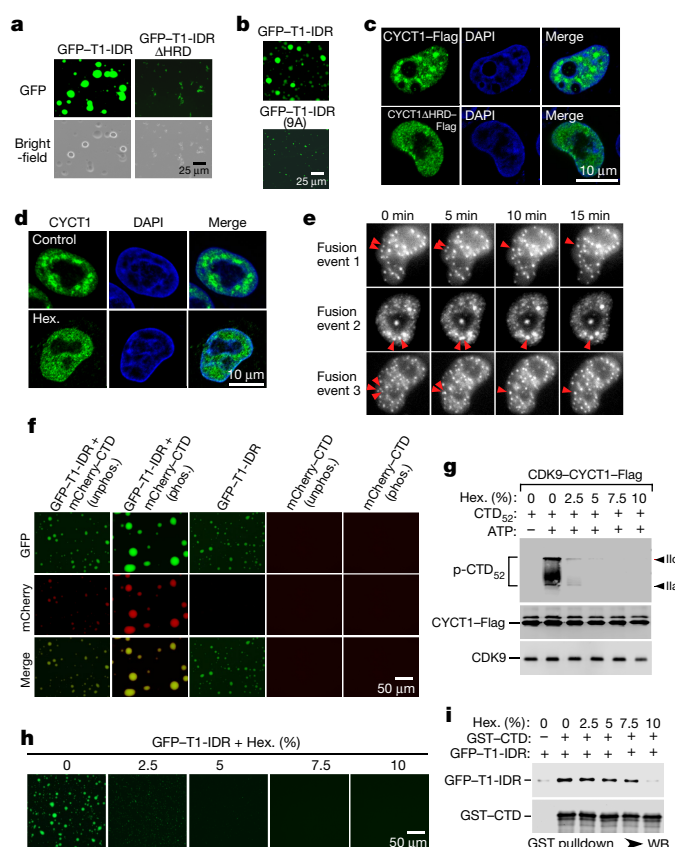
been attributed to the localization of P-TEFb to the nuclear speckles enriched with transcription and splicing factors<sup>20,21</sup>. Importantly, HRD and its histidine cluster were needed to target CYCT1 to the speckles (Fig. 4c and Extended Data Figs. 5h, 7e), and the same has also been reported for DYRK1A<sup>22</sup>. Mirroring its disruption of the GFP-T1-IDR droplets in vitro, 1,6-hexanediol also quickly disassembled the CYCT1 nuclear speckles (Fig. 4d). Finally, when performing time-lapse phase-contrast imaging of cells expressing eGFP-CYCT1, we observed multiple cells in which CYCT1 speckles displayed at least one fusion event within the period of a few minutes, demonstrating their dynamic and liquid-like properties (Fig. 4e).

Following previous findings<sup>23,24</sup>, we asked whether the HRD-containing IDR could recruit the CTD into phase-separated droplets, given their demonstrated direct interaction. Indeed, although the CTD itself is a low-complexity sequence, recombinant mCherry-CTD<sup>23</sup> (Extended Data Fig. 5d) alone did not phase separate, but it was readily incorporated into droplets when incubated together with GFP-T1-IDR (Fig. 4f).

During transcription, the CTD is phosphorylated first by CAK in TFIIF and then by P-TEFb<sup>1,4,5</sup>. Pre-phosphorylation by CAK (Extended Data Fig. 8c) not only enhanced the incorporation of mCherry-CTD into the GFP-T1-IDR droplets, but also promoted phase separation overall by producing bigger and brighter droplets (Fig. 4f). Consistently, the pre-phosphorylated GST-CTD precipitated more GFP-T1-IDR (Extended Data Fig. 8d). Underscoring the physiological relevance of these observations, hyperphosphorylated Pol IIo is known to preferentially localize in the nuclear speckles<sup>25</sup>.

Although both CTD hyperphosphorylation by CDK9 and HRD-mediated droplet formation were largely inhibited by 2.5% 1,6-hexanediol, the HRD-CTD binding (and CDK9-CYCT1 interaction) was not substantially inhibited until 1,6-hexanediol was at 10% (Fig. 4g–i). This important difference suggests that phase separation, which is caused by weak, multivalent interactions among the HRDs and is easily disrupted by 1,6-hexanediol, is critical for the hyperphosphorylation of the CTD.





**Fig. 4 | The hyperphosphorylation of CTD by P-TEFb is promoted by CYCT1 IDR, which forms phase-separated droplets and/or speckles in an HRD-dependent manner and recruits the CTD into these compartments.** **a, b,** Solutions containing the indicated proteins at 6 mg ml<sup>-1</sup> (**a**) or 2 mg ml<sup>-1</sup> (**b**) were examined with a microscope for fluorescence (GFP) or under white light (bright field). **c,** HeLa cells expressing wild-type CYCT1-Flag or CYCT1ΔHRD-Flag were examined by indirect immunofluorescence staining with anti-Flag monoclonal antibody. DNA was counterstained using DAPI. **d,** HeLa cells were treated with 10% 1,6-hexanediol (Hex.) or without 1,6-hexanediol (control) for 1 min and then analysed by immunofluorescence staining with anti-CYCT1 antibody. **e,** HeLa cells expressing eGFP-CYCT1 were examined by time-lapse phase-contrast imaging. Three separate cell nuclei are shown, in which CYCT1 speckles underwent spontaneous fusions as indicated by the arrows. **f,** Solutions containing GFP-T1-IDR at 3 mg ml<sup>-1</sup> and/or phosphorylated (phos.) or unphosphorylated (unphos.) mCherry-CTD at 1.2 mg ml<sup>-1</sup> were examined under a fluorescence microscope as in **a, g**. Kinase reactions as in Fig. 1a were performed in the presence of the indicated concentrations of 1,6-hexanediol. p-CTD<sub>52</sub> was detected by western blotting. Levels of CYCT1-Flag and its bound CDK9 in each reaction are also shown. **h,** 1,6-Hexanediol was present at the indicated concentrations in solutions containing 6 mg ml<sup>-1</sup> wild-type GFP-T1-IDR and 37.5 mM NaCl, which were examined under a fluorescence microscope. **i,** Immobilized GST-CTD was incubated with recombinant GFP-T1-IDR in reactions containing the indicated concentrations of 1,6-hexanediol. The bait and bound proteins were analysed by western blotting. All western blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.

On the other hand, the relatively drug-resistant HRD-CTD binding is probably key to the recruitment of the CTD to CDK9 in droplets, but is insufficient to establish the optimal environment for hyperphosphorylation (Extended Data Fig. 9).

In summary, our data indicate that in at least two kinases—CYCT1 of P-TEFb and DYRK1A—the HRD, a low-complexity domain of previously unknown function, promotes the hyperphosphorylation of the CTD by targeting the CTD as well as inducing phase separation in vitro and in cells. The phase-separated droplets and speckles

compartmentalize the kinase and substrate to enable highly efficient reactions, which results in the hyperphosphorylation of the CTD and robust transcriptional elongation and RNA processing.

FUS and TAF15, which are two proteins that contain low-complexity domain and that are active in transcription initiation, have previously been shown to form phase-separated hydrogels that trap the CTD<sup>23,24</sup>. Additionally, the bidirectionally transcribed enhancers and resulting antisense transcripts have been proposed to control initiation in part through phase separation<sup>26</sup>. What was unknown until now is whether any transcription factors and the CTD are involved in droplet formation after initiation<sup>1</sup>. Our finding that phase separation is induced by CYCT1 of P-TEFb (a well-defined transcription elongation factor) and DYRK1A (a probable gene-specific elongation factor<sup>6</sup>) has expanded the regulatory roles of phase separation to the next stage of the transcription cycle. Furthermore, these studies show that some key initiation and elongation factors that phase separate should no longer be viewed as passive passengers waiting to be picked up by the CTD. Rather, they have active roles in recruiting Pol II through multivalent interactions to their droplets and/or speckles that function as hubs where much of transcription and RNA processing is dynamically controlled.

## Data availability

Uncropped scans of all western blots are provided in Supplementary Figure 1. The raw slowSPT and spaSPT data are freely available in Spot-On readable CSV and Matlab formats in the form of single-molecule trajectories at Zenodo (<https://zenodo.org/record/1215836>). The Spot-On Matlab code is available, together with a step-by-step guide, at Gitlab (<https://gitlab.com/tjian-darzacq-lab/spot-on-matlab>). All other data are available from the corresponding author on reasonable request.

## Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0174-3>.

Received: 11 November 2017; Accepted: 2 May 2018;

Published online: 30 May 2018

- Harlen, K. M. & Churchman, L. S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat. Rev. Mol. Cell Biol.* **18**, 263–273 (2017).
- Eick, D. & Geyer, M. The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem. Rev.* **113**, 8456–8490 (2013).
- Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: new twists in the tail. *Nat. Struct. Mol. Biol.* **23**, 771–777 (2016).
- Zhou, Q., Li, T. & Price, D. H. RNA polymerase II elongation control. *Annu. Rev. Biochem.* **81**, 119–143 (2012).
- Kwak, H. & Lis, J. T. Control of transcriptional elongation. *Annu. Rev. Genet.* **47**, 483–508 (2013).
- Di Vona, C. et al. Chromatin-wide profiling of DYRK1A reveals a role as a gene-specific RNA polymerase II CTD kinase. *Mol. Cell* **57**, 506–520 (2015).
- Taube, R., Lin, X., Irwin, D., Fujinaga, K. & Peterlin, B. M. Interaction between P-TEFb and the C-terminal domain of RNA polymerase II activates transcriptional elongation from sites upstream or downstream of target genes. *Mol. Cell. Biol.* **22**, 321–331 (2002).
- Darzacq, X. et al. In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **14**, 796–806 (2007).
- Hansen, A. S., Pustova, I., Catoggio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* **6**, e25776 (2017).
- Hansen, A. S. et al. Robust model-based analysis of single-particle tracking experiments with Spot-On. *eLife* **7**, e33125 (2018).
- Xiang, J. et al. DYRK1A regulates Hap1-Dcaf7/WD68 binding with implication for delayed growth in Down syndrome. *Proc. Natl Acad. Sci. USA* **114**, E1224–E1233 (2017).
- Dosztányi, Z., Csizmek, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).
- Mitrea, D. M. & Kriwacki, R. W. Phase separation in biology: functional organization of a higher order. *Cell Commun. Signal.* **14**, 1 (2016).
- Courchaine, E. M., Lu, A. & Neugebauer, K. M. Droplet organelles? *EMBO J.* **35**, 1603–1612 (2016).
- Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

16. Lin, Y., Protter, D. S., Rosen, M. K. & Parker, R. Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Mol. Cell* **60**, 208–219 (2015).
17. Ying, Y. et al. Splicing activation by Rbfox requires self-aggregation through its tyrosine-rich domain. *Cell* **170**, 312–323.e310 (2017).
18. Strom, A. R. et al. Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017).
19. Molliex, A. et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015).
20. Herrmann, C. H. & Mancini, M. A. The Cdk9 and cyclin T subunits of TAK/P-TEFb localize to splicing factor-rich nuclear speckle regions. *J. Cell Sci.* **114**, 1491–1503 (2001).
21. Marcello, A. et al. Recruitment of human cyclin T1 to nuclear bodies through direct interaction with the PML protein. *EMBO J.* **22**, 2156–2166 (2003).
22. Salichs, E., Ledda, A., Mularoni, L., Albà, M. M. & de la Luna, S. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet.* **5**, e1000397 (2009).
23. Kato, M. et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).
24. Kwon, I. et al. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell* **155**, 1049–1060 (2013).
25. Mortillaro, M. J. et al. A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. *Proc. Natl Acad. Sci. USA* **93**, 8253–8257 (1996).
26. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).

**Acknowledgements** We thank S. McKnight, M. Geyer, J. Hurley and their colleagues for providing the various expression plasmids, and U. Schulze-Gahmen for technical help. This work was supported by the National Institutes of Health grant R01AI041757 to Q.Z. and the California Institute of Regenerative Medicine grant LA1-08013 to X.D.

**Reviewer information** *Nature* thanks J. Lis, D. Taatjes and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** H.L., X.D. and Q.Z. conceived the studies. H.L., D.Y. and R.L. performed kinase reaction assays, cell culture, immunofluorescence staining and droplet formation experiments and analyses. A.S.H. performed and analysed the single-particle tracking experiment. S.G. performed and analysed the FRAP assay. H.L. and A.H. performed and analysed the time-lapse phase-contrast imaging experiment. H.L., A.S.H. and Q.Z. wrote the manuscript, and all authors contributed ideas and reviewed the manuscript.

**Competing interests** The authors declare no competing interests.

#### Additional information

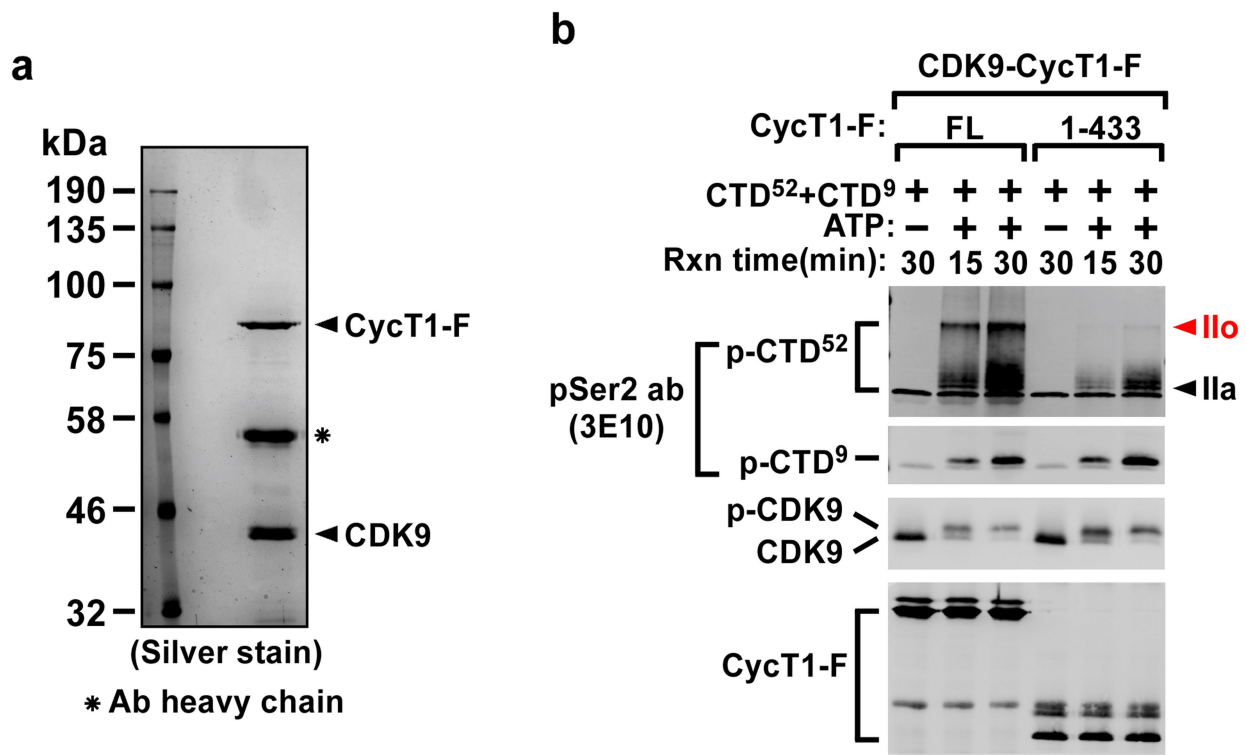
**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0174-3>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0174-3>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

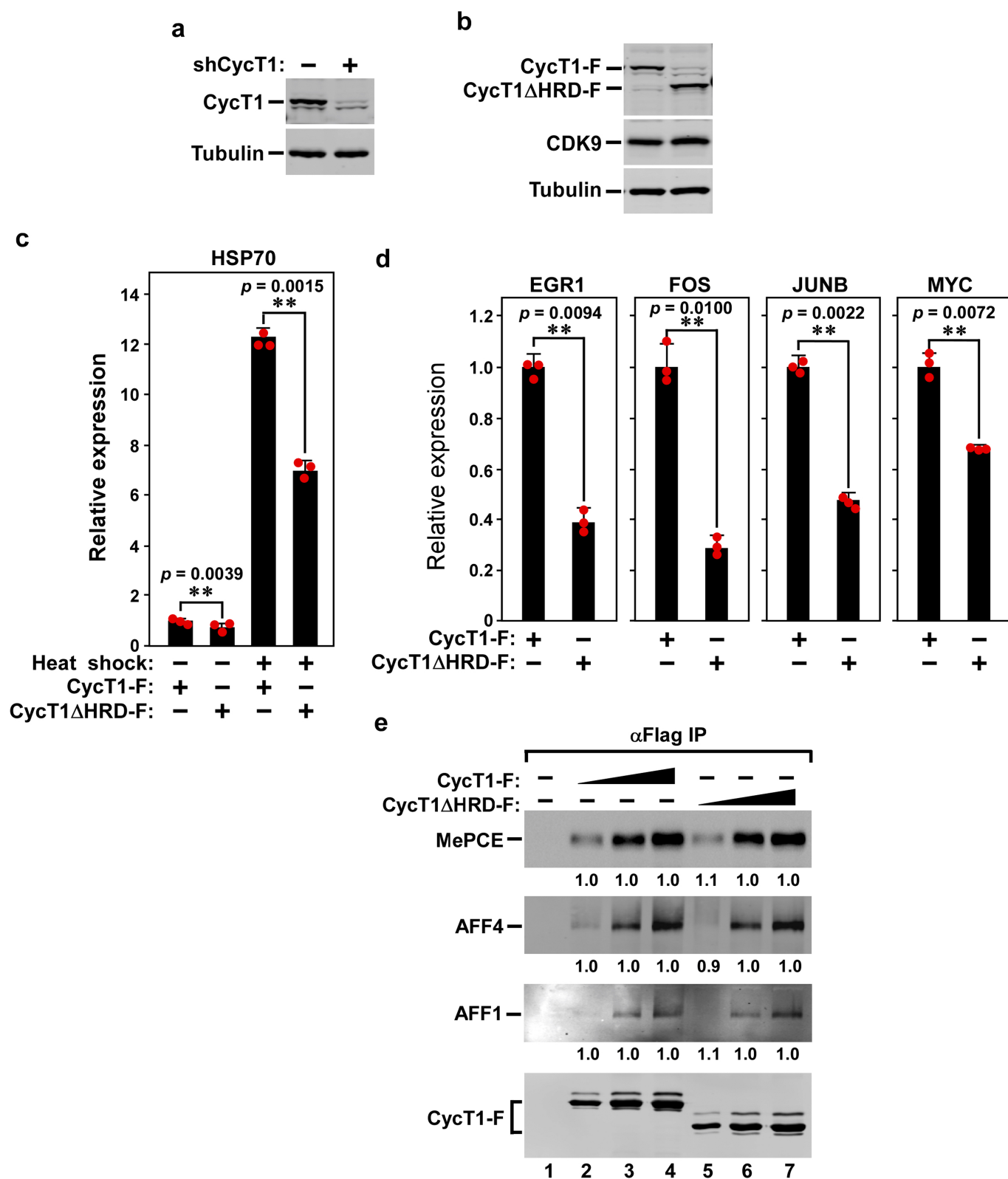
**Correspondence and requests for materials** should be addressed to Q.Z.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Extended Data Fig. 1 | The CYCT1 HRD-dependent hyperphosphorylation of Pol II CTD<sub>52</sub> by affinity-purified CDK9–CYCT1–Flag dimer can also be detected with anti-phospho-Ser2 antibody. a**, Examination of the affinity-purified CDK9–CYCT1–Flag dimer containing wild-type CYCT1–Flag by SDS–PAGE and silver staining. The anti-Flag affinity-purification was performed under high salt plus detergent (1 M KCl + 1% NP-40) conditions to strip away all the P-TEFb-associated factors but keep the CDK9–CYCT1 interaction intact. **b**, Affinity-purified CDK9–CYCT1–Flag heterodimers containing

the indicated CYCT1–Flag proteins were tested in kinase reactions containing a mixture of GST-fused CTD<sub>52</sub> and CTD<sub>9</sub> as the substrates. The phosphorylated p-CTD<sub>52</sub> and p-CTD<sub>9</sub> were detected by western blotting with the anti-phospho-Ser2 antibody 3E10. Although a very similar pattern of CTD phosphorylation was detected with both anti-phospho-Ser2 and anti-phospho-Ser5 antibodies, the unphosphorylated CTD present in the ATP(–) lanes was only detected by the former antibody, making the phospho-Ser5 antibody a preferred choice for detecting CTD phosphorylation in these kinase reactions.



Extended Data Fig. 2 | See next page for caption.



**Extended Data Fig. 2 | The CYCT1 HRD is required for efficient transcription of human immediate early genes as well as the *HSP70-1* gene under both basal and heat-shock conditions.** **a**, Confirmation by western blotting of the knockdown of endogenous CYCT1 expression in HEK293T cells expressing the *CYCT1* (also known as *CCNT1*)-specific shRNA (shCYCT1). **b**, Anti-Flag western blotting analysis of the expression of either wild-type CYCT1-Flag or CYCT1 $\Delta$ HRD-Flag from the shCYCT1-resistant plasmid introduced into the knockdown cells. The  $\alpha$ -tubulin and CDK9 levels were used as controls. **c**, The HRD-dependent transcription of four cellular immediate early genes. The mRNA levels of the indicated immediate early genes in the knockdown cells expressing wild-type CYCT1 or CYCT1 $\Delta$ HRD-Flag (analysed in **b**) were examined by qRT-PCR, normalized to that of *GAPDH* and shown. The activity in the first column of each group was set to 1. Data are mean  $\pm$  s.d.,  $n = 3$ , and  $P$  values from two-tailed Student's  $t$ -test. **d**, The CYCT1 HRD is

required for optimal *HSP70-1* transcription under both basal and heat-shock conditions. The *HSP70* mRNA levels in the knockdown cells expressing wild-type CYCT1 or CYCT1 $\Delta$ HRD were examined under heat-shock or non-heat-shock conditions, as in **c**. **e**, HeLa cells were transfected in twofold increments with the plasmid expressing either wild-type CYCT1-Flag or CYCT1 $\Delta$ HRD-Flag. Anti-Flag immunoprecipitates (IP) from nuclear extracts were examined by western blotting for the proteins labelled on the left. The levels of the co-precipitated MEPCE, AFF4 and AFF1 were first quantified and then normalized to that of their corresponding CYCT1-Flag bait. The levels of MEPCE, AFF4 and AFF1 bound to low, middle and high levels of CYCT1 $\Delta$ HRD-Flag were then divided by those of MEPCE, AFF4 and AFF1 bound to the corresponding levels of wild-type CYCT1-Flag and shown in lanes 5–7, with the numbers in lanes 2–4 set to 1 as a reference.

**a**

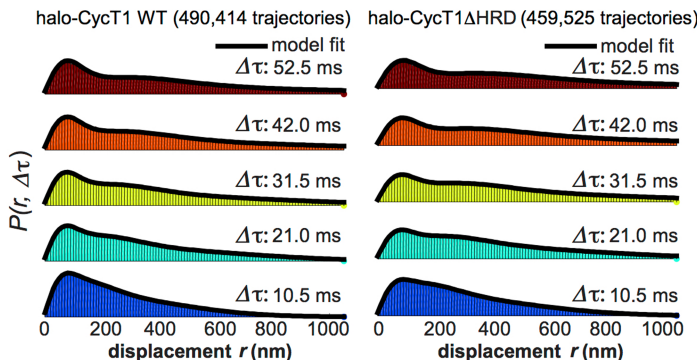
Estimates of parameters extracted from fitting of FRAP data (with  $\pm$  standard errors) wherever applicable

	$D_{app}$ ( $\mu m^2 sec^{-1}$ )	$k_{on}^*$ ( $sec^{-1}$ )	$k_{off}$ ( $sec^{-1}$ )	Sum of residuals
halo-CycT1 WT	$1.66 \pm 0.08$	$0.0044 \pm 0.0007$	$0.0176 \pm 0.0019$	0.00173
halo-CycT1 $\Delta$ HRD	$4.15 \pm 0.9$	$0.0657 \pm 0.04$	$0.231 \pm 0.07$	0.00871

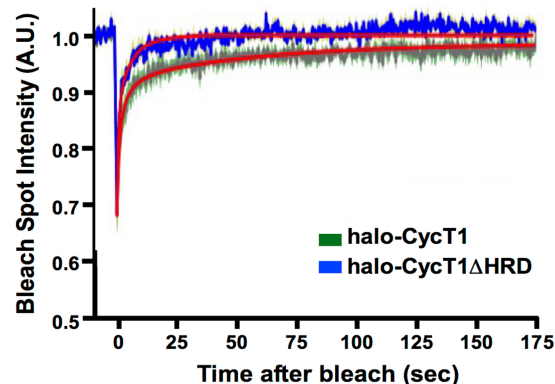
Parameters evaluated from the fitted FRAP data above with the corresponding ranges

	$D_{app}$ ( $\mu m^2 sec^{-1}$ )	$C_{eq}$	$\tau_{on}$ (sec)	$\tau_{off}$ (sec)
halo-CycT1 WT	$1.66 \pm 0.08$	$\sim 0.2$	226 (196-270)	56 (51-64)
halo-CycT1 $\Delta$ HRD	$4.15 \pm 0.9$	$\sim 0.2$	15 (9-38)	4 (3-6)

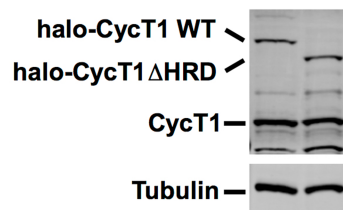
**d**



**b**



**c**

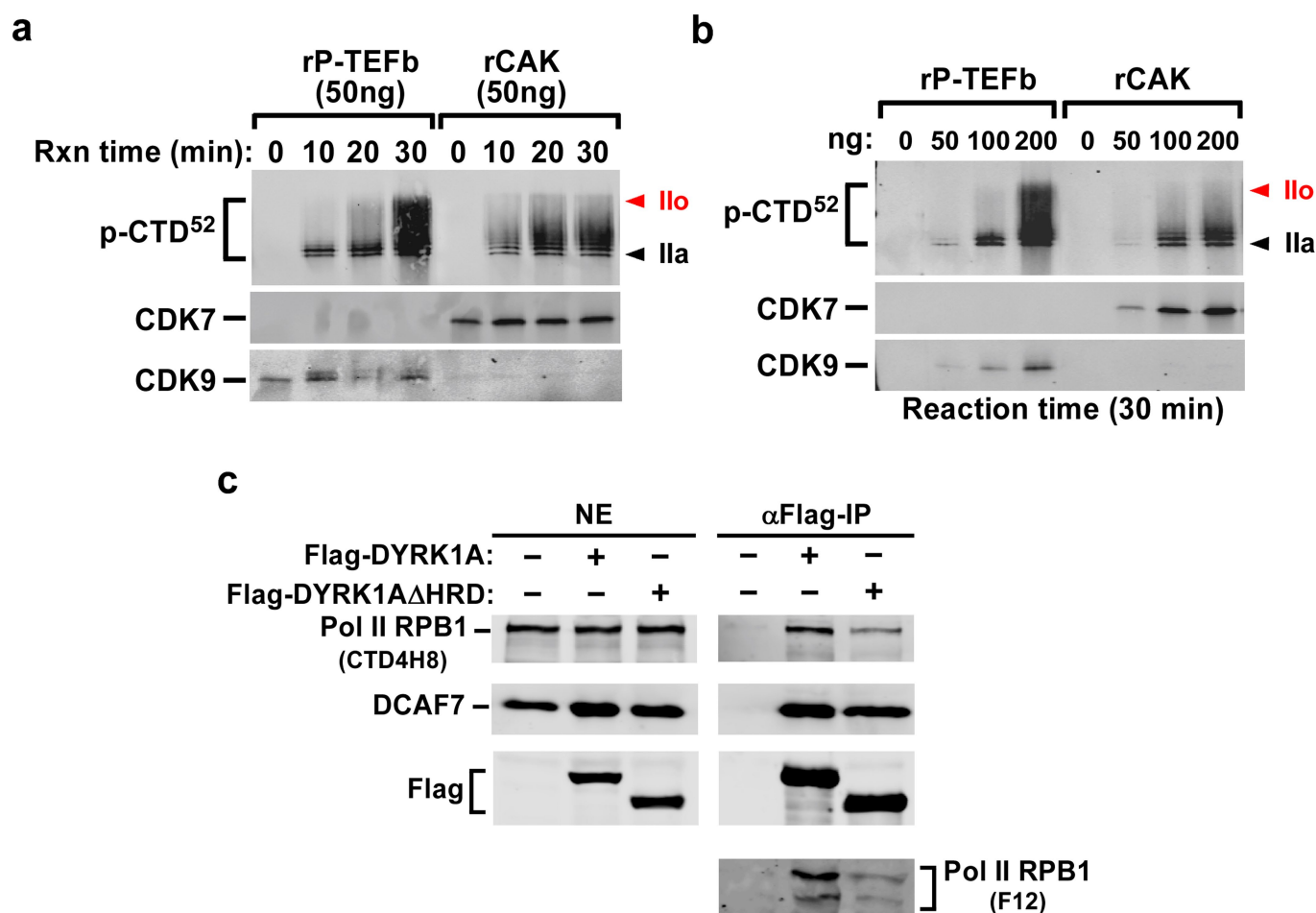


**e**

3-state Spot-On model fit (1 bound, 2 free states)			
U2OS halo-CycT1	WT	$\Delta$ HRD	
$F_{BOUND}$	$13.0\% \pm 2.0\%$	$8.9\% \pm 0.9\%$	
$D_{BOUND}$ ( $\mu m^2/s$ )	$0.04 \pm 0.0$	$0.04 \pm 0.0$	
$F_{SLOW}$	$50.2\% \pm 4.3\%$	$41.5\% \pm 9.5\%$	
$D_{SLOW}$ ( $\mu m^2/s$ )	$1.18 \pm 0.10$	$1.39 \pm 0.16$	
$F_{FAST}$	$39.8\% \pm 4.0\%$	$49.6\% \pm 10.4\%$	
$D_{FAST}$ ( $\mu m^2/s$ )	$5.70 \pm 0.09$	$6.83 \pm 1.01$	

**Extended Data Fig. 3 | The HRD promotes the binding of CYCT1 to activated gene expression array, endogenous genes and the RNA Pol II CTD.** **a**, Estimates of parameters extracted from fitting of FRAP data (top) and parameters calculated from the fitted FRAP data (bottom). **b**, Model fit overlaid on raw FRAP data. A full description of how the reaction-diffusion FRAP model was fitted to the data is provided in Supplementary Information. **c**, Anti-CYC1 western blotting analysis of the levels of endogenous CYCT1, and the stably expressed Halo-CYC1

and Halo-CYC1 $\Delta$ HRD, in lysates of the engineered U2OS cell lines. The  $\alpha$ -tubulin levels provided an internal control. **d**, Histograms of displacements at the indicated  $\Delta\tau$  with three-state model fit overlaid. The three-state model is described in Supplementary Information. **e**, Table of best-fit parameters from fitting Spot-On to the raw displacements for three independent replicates (about 5–10 cells per replicate). Values in the table are mean  $\pm$  s.d.,  $n = 3$ .

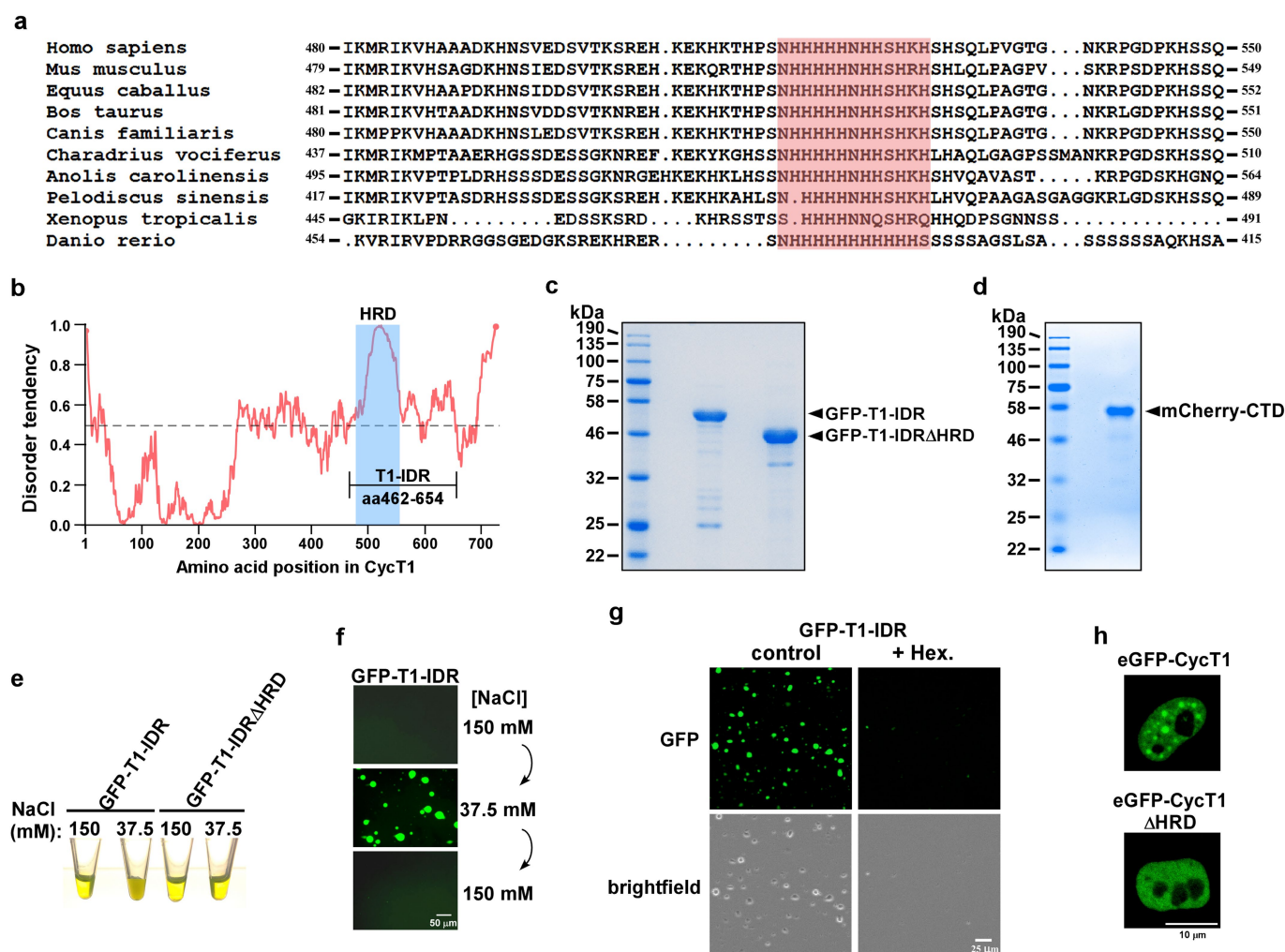


**Extended Data Fig. 4 | Examination of recombinant P-TEFb and CAK for their phosphorylation of the Pol II CTD in kinase reactions and contribution of the DYRK1A HRD to DYRK1A–Pol II interaction.**

**a, b,** Compared to CDK9 in recombinant P-TEFb, CDK7 in recombinant CAK (CDK7–CYCH–MAT1) shows decreased ability to hyperphosphorylate CTD<sub>52</sub> in a time- and dosage-dependent manner. The indicated amounts of baculovirus-produced recombinant P-TEFb or CAK (Millipore) were added to in vitro kinase reactions that also

contained GST–CTD<sub>52</sub> as the substrate. The reactions were performed for the indicated periods of time. The products were analysed by western blotting with the phospho-Ser5 antibody. CDK7 in recombinant CAK and CDK9 in recombinant P-TEFb were also examined by western blotting.

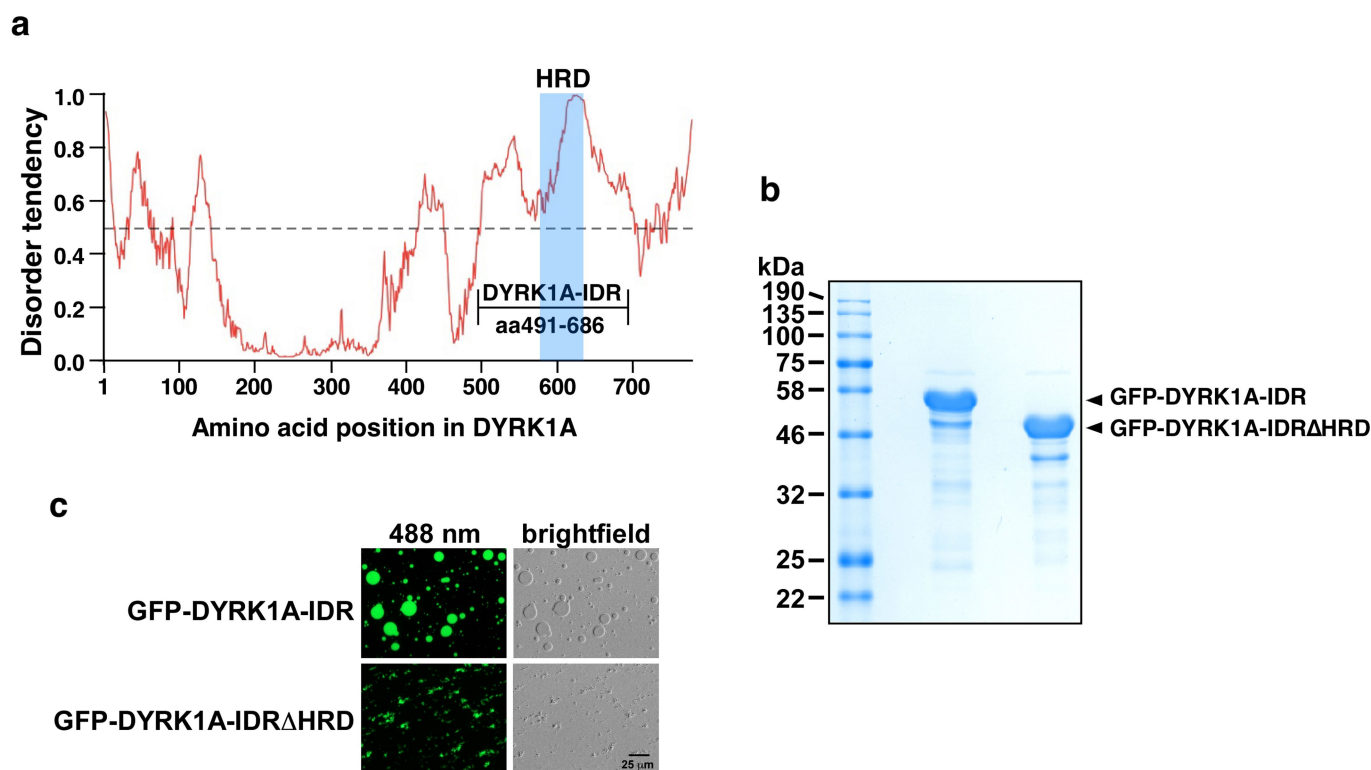
**c,** The deletion of the HRD causes DYRK1A to decrease interaction with RNA Pol II but not DCAF7. Nuclear extract (NE) of HeLa cells expressing the indicated proteins and anti-Flag immunoprecipitates derived from NE were analysed by western blotting with the various antibodies as labelled.



**Extended Data Fig. 5 | GFP-T1-IDR purified from recombinant *E. coli* forms phase-separated liquid droplets that are HRD-dependent and highly sensitive to elevated salt concentration and exposure to 1,6-hexanediol. a**, Alignment of CYCT1 amino acid sequences in the HRD regions among the indicated vertebrate species, with the central histidine cluster shaded in red. **b**, Intrinsic disorder tendency was predicted by IUPred across the entire length of CYCT1. The scores are assigned between 0 and 1, and a score above 0.5 indicates disorder. The HRD region is shaded in blue. The longest stretch of IDR is labelled as T1-IDR, and its boundaries are marked. **c**, **d**, The C-terminally Strep-tagged GFP-T1-IDR fusions (**c**) and N-terminally His-tagged mCherry-CTD fusion (**d**) were purified from recombinant *E. coli* BL21

cells (Supplementary Information). Ten micrograms of each of the fusion proteins was examined by SDS-PAGE followed by Coomassie blue staining. **e**, Protein solutions containing either wild-type GFP-T1-IDR or GFP-T1-IDR $\Delta$ HRD at 6 mg ml<sup>-1</sup> were adjusted to the indicated salt concentrations and their appearances in Eppendorf tubes are shown. **f**, NaCl concentrations in the wild-type GFP-T1-IDR solution were changed in the indicated sequential order and then examined under a fluorescence microscope. **g**, Protein solutions containing wild-type GFP-T1-IDR at 6 mg ml<sup>-1</sup> were adjusted to 37.5 mM NaCl with or without 10% 1,6-hexanediol, and then examined under a fluorescence microscope. **h**, Live-cell images of a HeLa cell expressing either wild-type eGFP-CYC1 or eGFP-CYC1 $\Delta$ HRD at levels similar to that of endogenous CYCT1.

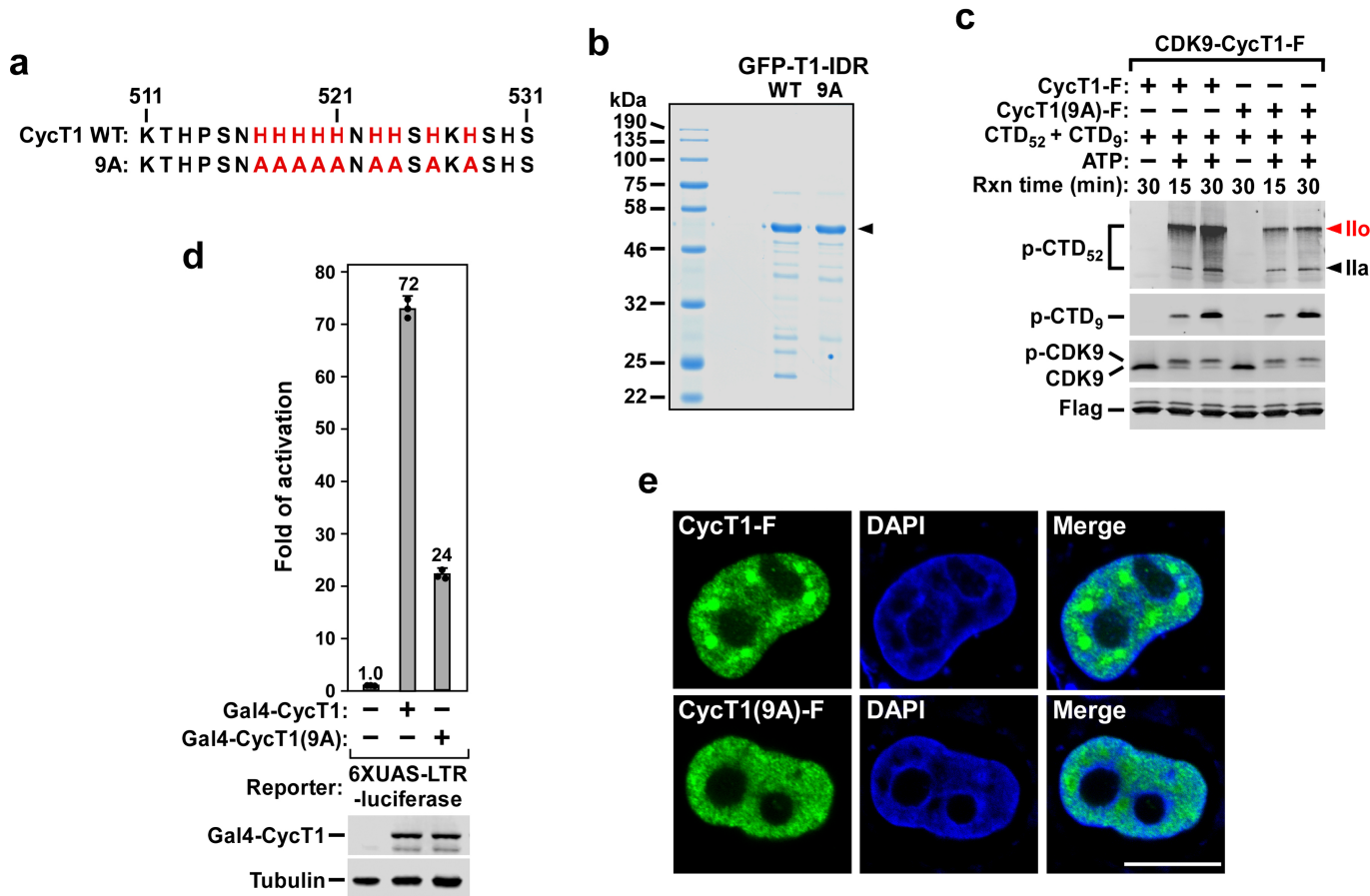




**Extended Data Fig. 6 | The longest stretch of IDR in DYRK1A promotes formation of phase-separated droplets in an HRD-dependent manner.**

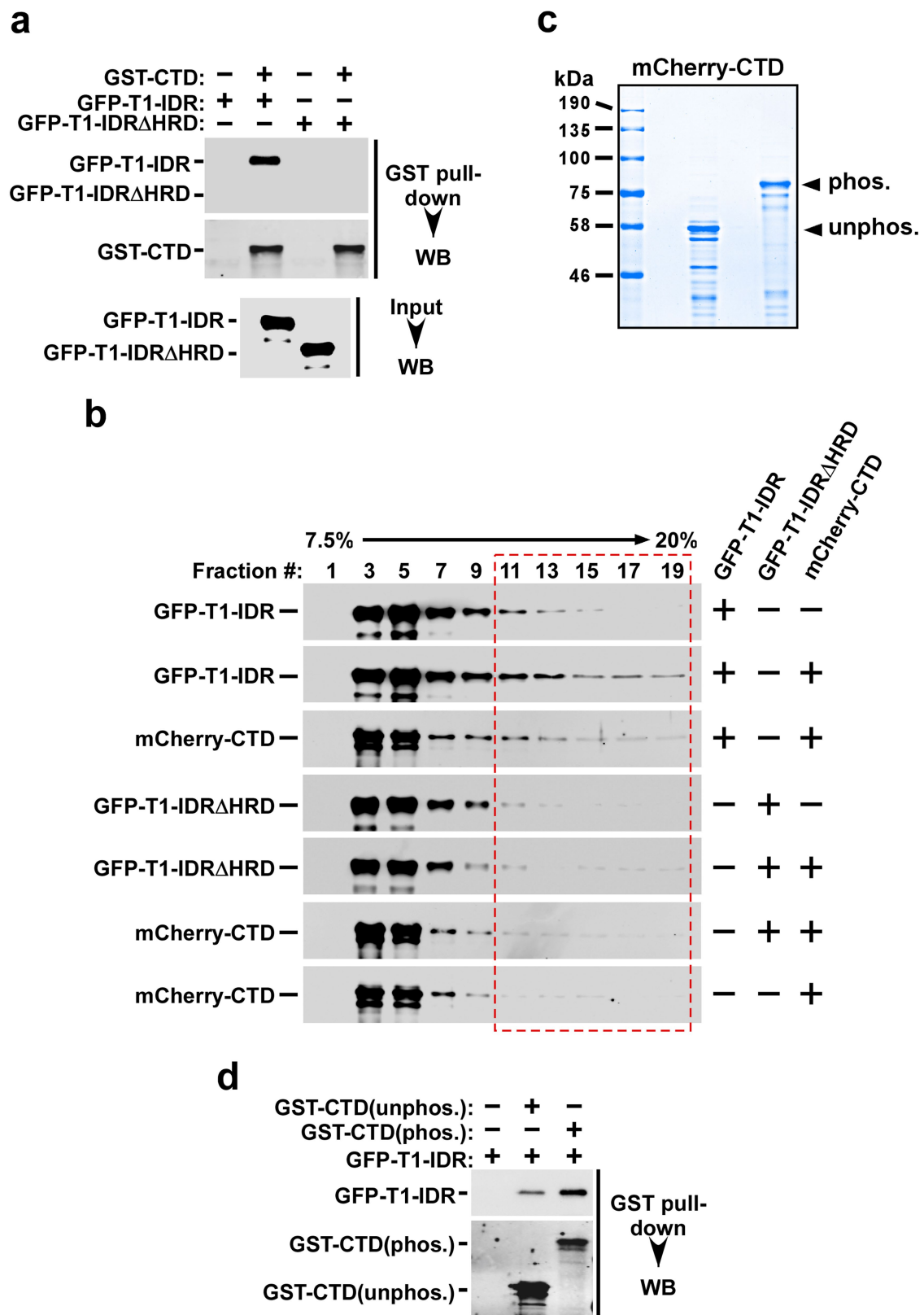
**a**, Intrinsic disorder tendency was predicted by IUPred across the entire length of DYRK1A. The scores are assigned between 0 and 1, and a score above 0.5 indicates disorder. The HRD region is shaded in blue. The longest stretch of IDR is labelled as DYRK1A-IDR and its boundaries are marked. **b**, The indicated C-terminally Strep-tagged GFP fusions were

purified from recombinant *E. coli* BL21 cells. Two micrograms each of the fusion proteins was examined by SDS-PAGE followed by Coomassie blue staining. **c**, Solutions containing  $5 \text{ mg ml}^{-1}$  of the indicated fusion proteins,  $37.5 \text{ mM NaCl}$  and  $10\% \text{ PEG8000}$  were trapped between coverslips and examined with a microscope under either fluorescent (488 nm) or normal white light (brightfield).



**Extended Data Fig. 7 | The central histidine cluster within the CYCT1 HRD is essential for promotion of phase separation by CYCT1 IDR in vitro and in cells, and for P-TEFb to hyperphosphorylate the Pol II CTD and activate HIV transcription.** **a**, The nine histidines in the central histidine cluster within the CYCT1 HRD are highlighted in red and changed to alanines in the 9A mutant. **b**, The C-terminally Strep-tagged GFP-T1-IDR fusions containing either wild-type GFP-T1-IDR or the GFP-T1-IDR(9A) mutant sequence were purified from *E. coli* BL21 cells and examined by SDS-PAGE followed by Coomassie blue staining. **c**, CDK9-CYC1-Flag heterodimers containing the indicated CYCT1-Flag proteins were affinity-purified from HeLa cells and tested

in kinase reactions, with the reaction products analysed as in Fig. 1a. **d**, Plasmids expressing the Gal4 DNA binding domain fused to the indicated CYCT1 proteins were co-transfected into HeLa cell with a HIV-1 LTR-luciferase reporter construct containing the UAS for Gal4. Luciferase activities in cell extracts were measured and analysed as in Fig. 1b. **e**, Fixed and permeabilized HeLa cells expressing wild-type CYCT1-Flag or CYCT1(9A)-Flag were examined by indirect immunofluorescence with the mouse anti-Flag monoclonal antibody and Alexa Fluor 488-conjugated goat anti-mouse secondary antibody. DNA was counterstained using DAPI.

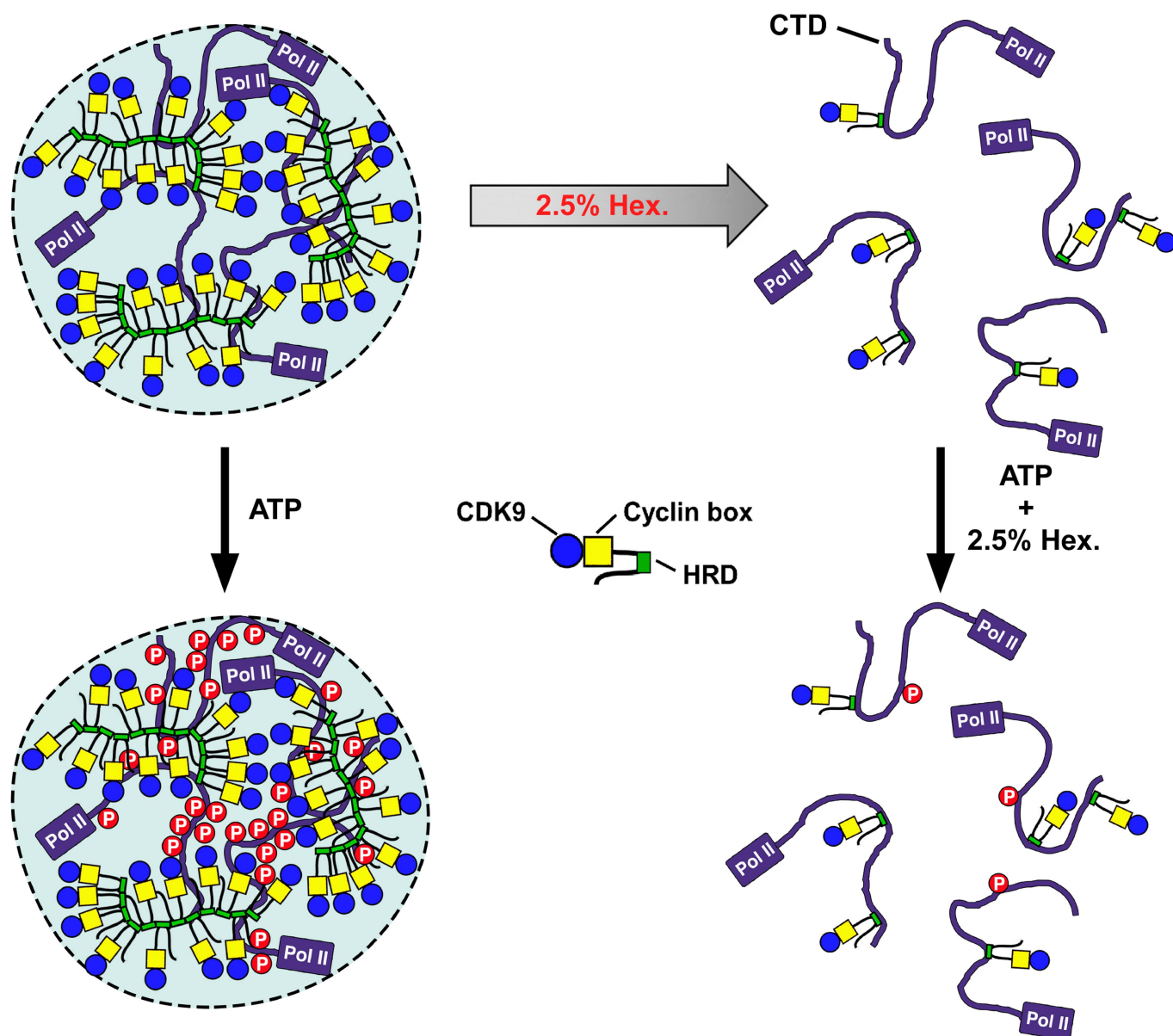


Extended Data Fig. 8 | See next page for caption.

**Extended Data Fig. 8 | CYCT1 binds directly to the Pol II CTD in an HRD-dependent manner and the binding is enhanced after the CTD is phosphorylated by CAK (CDK7–CYCH–MAT1).** **a**, Immobilized GST–CTD was incubated with recombinant GFP–T1-IDR or GFP–T1-IDR $\Delta$ HRD. The input (2.5%) and the bound proteins were analysed by western blotting. **b**, Binding reactions containing the purified recombinant fusion proteins indicated on the right were analysed in a 7.5 to 20% glycerol gradient containing 500 mM NaCl plus 0.5% NP-40, which was centrifuged at 55,000 r.p.m. and 4 °C for 13 h. The indicated fractions were analysed by western blotting to detect the distributions of proteins marked

on the left. The entire length of the CTD could be bound by varying numbers of IDRs, resulting in the formation of a series of complexes with broad distributions in the gradient. **c**, mCherry–CTD was incubated with or without immobilized CAK for 6 h in kinase reactions and then analysed by SDS–PAGE and Coomassie blue staining. **d**, Immobilized GST–CTD was incubated with (phos.) or without (unphos.) CAK for 6 h in kinase reactions. After washing, the GST–CTD beads were incubated with GFP–T1-IDR. The indicated proteins were eluted off the beads and analysed by western blotting.





**Extended Data Fig. 9 | A model depicting how P-TEFb uses the CYCT1 HRD to target and recruit the Pol II CTD into a phase-separated compartment—which is formed by weak, multivalent homotypic interactions among the HRDs—to enable highly efficient**

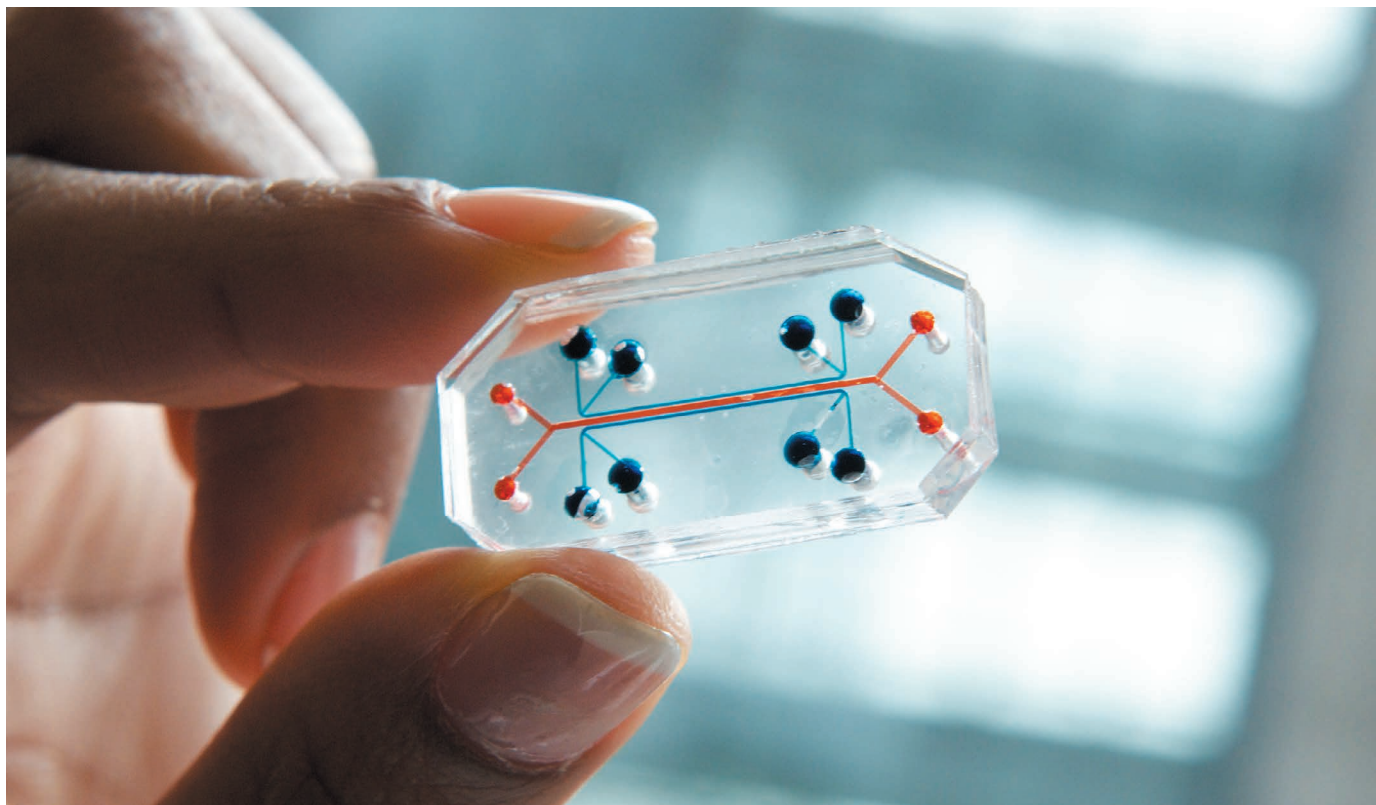
**phosphorylation of CTD by P-TEFb in the presence of ATP. At 2.5% 1,6-hexanediol, the HRD-mediated phase separation but not the direct HRD-CTD binding is disrupted, making it impossible for P-TEFb to hyperphosphorylate the CTD.**

## TECHNOLOGY FEATURE

# THE THIRD DIMENSION IN STEM-CELL CULTURE

*Human stem cells are yielding 3D miniature tissues that can be used to study both normal human biology and human disease in vitro.*

DARPA/WYSS INSTITUTE



Microfluidic devices, such as this 'lung-on-a-chip', provide sophisticated tools for stem-cell biologists.

BY C. Y. TACHIBANA

Stem-cell researcher Toshiro Sato places a culture dish under the microscope in his laboratory at Keio University School of Medicine in Tokyo. What he sees is not a sheet of cells, but something more complex — delicate spheres of tissue that are barely visible to the naked eye. These are organoids, 3D structures that develop when stem cells are given the proper environmental conditions to differentiate and arrange themselves into forms with properties similar to those of particular tissues or organs.

Organoids can be derived from pluripotent stem cells, which have the potential to form any body tissue type — be it muscle, skin, gut or brain. When they are grown in the

appropriate conditions, for example with specific growth factors, the stem cells self-organize into structures in which the different cell types are arranged similarly to the *in vivo* tissue. By contrast, organs-on-chips (OOCs) are generally made by arranging stem cells and cells that have already differentiated into the required cell types on a microfabricated device in positions and structures appropriate for the planned experiments. The difference between organoids and OOCs, says developmental biologist Madeline Lancaster at the MRC Laboratory of Molecular Biology in Cambridge, UK, "is self-organization versus construction" (see 'Models in a nutshell').

Named 'Method of the Year' by *Nature Methods* last year, organoids are increasingly being used to study both normal development

and the progression of diseases. Sato's group, for instance, uses them to study the early stages of tumour formation. His team uses the CRISPR gene-editing technique to change the sequence of the DNA in the stem cells, and then see how that affects the development of the organoid. "It's a way to see causality," Sato says, "by testing if specific mutations recapitulate cancer development."

Both organoids and OOCs also have potential in assessing the efficacy and safety of drugs, chemicals and cosmetics, with possible applications in regenerative medicine. For instance, a challenge in randomized clinical trials is how to compare the effect of treatments because the genetics and life history of the participants can affect how they react, notes Donald Ingber, founding director ►

► of the Wyss Institute for Biologically Inspired Engineering at Harvard University in Cambridge, Massachusetts. OOCs and organoids derived from patients' cells can eliminate these confounding effects by creating intervention and control populations that have identical genetics and clinical history. However, the technologies are not without their challenges, such as how to scale up production to meet growing basic and applied research demand while maintaining the reproducibility and fidelity of the structure to the *in vivo* organs they represent.

### ORGANS IN A DISH

The pluripotent stem cells that researchers use to create organoids and OOCs include both naturally occurring embryonic stem (ES) cells and cells that are derived from differentiated cells such as fibroblasts and manipulated to revive their pluripotency, known as induced pluripotent stem cells (iPS cells). Lancaster, who uses organoids to study basic brain development and identify factors contributing to complex conditions such as autism and schizophrenia, says she uses ES cells for developing and testing protocols and new models.

Organoids used to study genetic disorders or for personalized medicine are usually made from adult stem cells or iPS cells that have been manipulated to be specific to the patient. Geneticist Hans Clevers and his group at the Hubrecht Institute in Utrecht, the Netherlands, use organoids derived from intestinal stem cells to predict how people with cystic fibrosis will respond to various medications. Cystic fibrosis can be caused by any of several mutations in a single gene, *CFTR*. Therapies exist, but they are expensive

and known to work only in patients with particular mutations. Clevers's group is now testing organoids derived from the 600 or so Dutch people with cystic fibrosis without any of those mutations. The rationale is that if the drug causes the person's organoids to swell under assay conditions, then the patient is likely to respond, too.

Similarly, Sato was part of the team that developed methods for generating intestinal organoids from adult stem cells<sup>1</sup>. He is now involved in a clinical trial to test whether tissue made from eight patients' own stem cells can safely be implanted into the gut to repair damage caused by the inflammatory disease ulcerative colitis.

Organoids are also proving their worth in cancer research and drug development. The international Human Cancer Models Initiative is developing 'next-generation' organoid models, in which certain DNA sequences have been annotated. When combined with clinical data, the organoids allow researchers to link their findings to patient characteristics and outcomes. Daniela S. Gerhard, director of the National Cancer Institute's Office of Cancer Genomics in Bethesda, Maryland, which is involved in the initiative, says that these organoids should become available from the American Type Culture Collection (ATCC) in Manassas, Virginia, later this year. About 150 different organoid models will be offered initially, and pricing has yet to be determined. The Hubrecht Organoid Technology biobank, of which Clevers is chief scientific officer, offers hundreds of different organoids derived from adult stem cells at €2,000–3,000 (US\$1,700–2,600) apiece. Clevers and his colleagues have used organoids derived from colorectal cancer cells from 18

patients to test 83 anticancer compounds. Patterns of drug resistance in the organoids corresponded with known drug-resistance mutations, suggesting that organoids could be used to predict a person's response to a particular drug<sup>2</sup>.

Organoids derived from ES cells and iPS cells are ideal for studying complex developmental processes, Clevers says, but they can take weeks or months to produce; those made from adult stem cells tend to require less time. The shorter process is also less likely to introduce variation, so these organoids tend to be more reproducible than those derived from ES or iPS cells.

The process of culturing organoids is similar to standard tissue-culture work. "New graduate students can easily grow them in a few weeks," says Sato. "It's easy to make something from stem cells," adds Lancaster; the difficulty lies in the interpretation. It's hard to be certain about which tissues are actually present in the resulting organoid.

### OOCs FOR CONSISTENCY

Sometimes called microphysiological systems or tissue chips, OOCs are based on structured microdevices called microfluidic chips, on which cells can be maintained in culture. Unlike organoids, which develop spontaneously from stem cells, an OOC is designed. The structure of the chip and the type and placement of cells determine the tissues that emerge and their arrangement. That means they are generally more consistent than organoids, says bioengineer Boyang Zhang at McMaster University in Hamilton, Canada. And OOCs can be more sophisticated, too, because developers can add engineered elements, such as sensors, 'vasculature' to promote fluid and gas exchange, and features that facilitate imaging, that cannot spontaneously arise in organoids.

Some tissue chips are available commercially. Bioengineers developing OOCs can get their starting chips made to order by commercial fabrication facilities and on-campus machine shops. A research team at the Wyss Institute has created 'hearts-on-chips' using microscope-slide coverslips coated with a synthetic polymer. Starting with iPS cells from people with Barth syndrome, a form of congenital cardiac disease, the team first coaxed the iPS cells to differentiate into heart-muscle cells (cardiomyocytes) and then grew those on a chip to produce tissue that they could test for function. The researchers were able to show that a mutation associated with Barth syndrome caused the cardiomyocytes to function abnormally. They were also able to correct the defect *in vitro*<sup>3</sup>.

TARA Biosystems in New York City, which was co-founded by Zhang, produces a cardiac OOC called Biowire for drug testing. TARA scientists take cardiomyocytes derived from iPS cells and place them in a microdevice containing a fine wire around which the cells grow. That helps the cardiomyocytes to

## Models in a nutshell

Organoids and organs-on-chips (OOCs) have applications in drug development, cosmetics testing, toxicology and personalized medicine. Here are some of their strengths and weaknesses.

#### Organoids

**Origin:** Generated from self-organizing embryonic stem cells (ES cells), induced pluripotent stem cells (iPS cells) or adult stem cells.

**Key strengths:** Studies of developmental processes as single cells give rise to organs, including in disease.

**Key challenges:** Increasing the reproducibility for applications that need a consistent outcome, such as drug testing, and finding ways to introduce or mimic vascularity.

**Learn more:** Cell Press webinar 'Organoids and beyond — 3D tissue in a dish' (see [go.nature.com/2jiumvb](http://go.nature.com/2jiumvb))

#### OOCs

**Tissue origin:** Typically, cell lines or differentiated cells isolated from a person, but can also be derived from stem cells.

**Key strengths:** Reproducibility and consistency. Can incorporate biomechanical features.

**Key challenges:** Using iPS cells and stem cells from patients rather than cell lines, to increase applicability to personalized medicine.

**Learn more:** National Academies of Sciences, Engineering, and Medicine webinar 'The NIH Microphysiological Systems Program' (see [go.nature.com/2hm3bch](http://go.nature.com/2hm3bch))



align in the regular arrangement required for function. The developing ‘muscle’ can then be electrically stimulated to become mature heart muscle that can contract and relax, mimicking the functioning of real heart tissue. Treatment with adrenaline, for example, increases contraction.

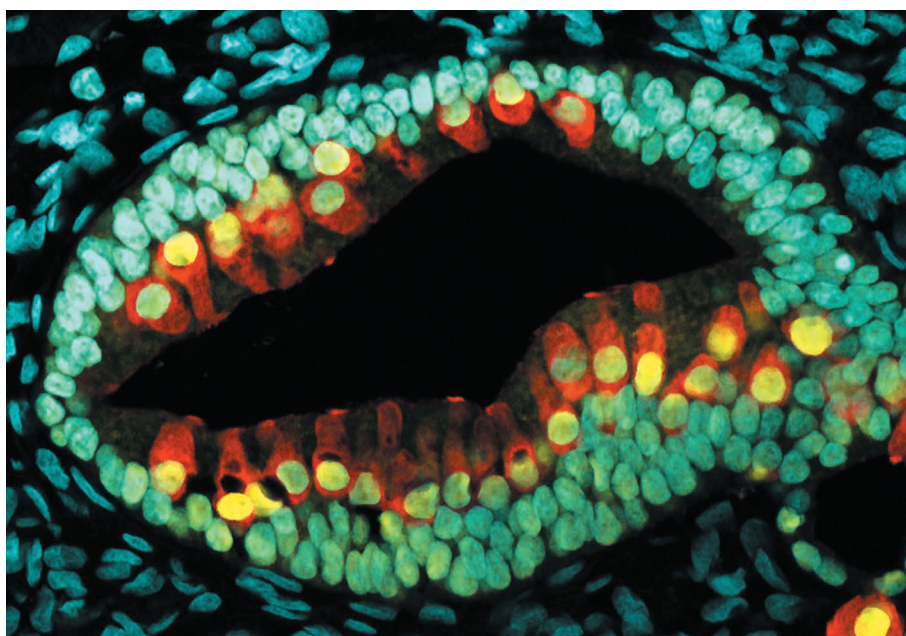
Nortis near Seattle, Washington, creates OOCs by solidifying a collagen matrix around a glass fibre, which is then removed to produce hollow channels that are several millimetres long and about 100 micrometres in diameter. Once the tubes are seeded with cells, tissues form within days. A single-use chip, which can run three parallel experiments, costs \$300.

The use of stem cells to generate tissue for OOCs will become more prominent because of their potential for regenerative and personalized medicine, says Nortis chief executive Thomas Neumann. But generating tissue from stem cells introduces issues if cells do not fully differentiate or mature, so quality assurance will become even more important, he says.

OOCs can be engineered to have levels of complexity that organoids generally cannot match. They can, for example, be made to mimic physiological properties such as tissue stretching, pulsation and peristalsis. Ingber launched the OOC field in his 2010 paper<sup>4</sup> on micromodels of the lung capillary–alveolar interface, coined ‘lungs-on-chips’. The chip has channels separated by a membrane that has alveolar cells on one side and vascular cells on the other. Breathing can be simulated by applying and releasing vacuum to side chambers. “The mechanical microenvironment is critical for getting *in vivo*-level function,” Ingber says. The team has used the simulated breathing system to test the toxicity of nanoparticles.

OOCs can also be linked together to create multi-organ ‘body-on-a-chip’ models. Michael Shuler and his bioengineering team at Cornell University in Ithaca, New York, have produced a single, closed OOC system with 14 chambers to represent organs with ‘barrier’ functions such as those seen in the lungs, and ‘non-barrier’ functions similar to those in the heart<sup>5</sup>.

Bioengineer Linda Griffith at the Massachusetts Institute of Technology in Cambridge and her team have built a ten-organ system in which material can flow from one organ to another, as it would in the body<sup>6</sup>. Some organs, such as the ‘brain’, originated from iPS cells; others, including the ‘liver’, came from cell lines or other differentiated cells. The tissues remained viable for up to four weeks, and the system displayed biological functionality. For example, a drug introduced into the gut tissue was passed to the liver, where it was metabolized. The project, Griffith says, involved a great deal of cost, effort and multidisciplinary coordination, requiring not only bioengineers and mechanical engineers, but also modellers to translate how the data



A cross-section of an inner-ear organoid.

from the chips apply to humans. In practice, Griffith says, scientists generally work with just two to four linked chips.

Developers of OOCs are now setting their sights on demonstrating the value and validity of the technology to industry and to regulatory agencies. Scaling up the use of OOCs will require manufacturing-friendly formats for production and high-throughput applications. One challenge for linked OOC systems is finding culture media and conditions that serve all tissue types — media that support liver cells are not always suitable for lung cells, for instance. Achieving accurate models also requires adjusting the numbers of cells and activities of the various chips so that they accurately represent how those organs would work in a full-sized human. But, in general, Ingber says, the field is becoming more user-friendly. “It’s getting to the point where it’s plug-and-play, and you don’t need to be a microsystems engineer,” he says.

#### WHAT’S NEXT?

As organoids and OOCs become used more widely, researchers are beginning to address more-sophisticated questions. Griffith and others are adding gut microbiomes to their platforms, for example in a model that links up gut, liver and brain to study Parkinson’s disease. She says that as costs go down and reproducibility improves, OOCs might begin to substitute for animals in experiments in which they are used as surrogates for humans. “We’re still in the early stage of thinking about biology in an engineering sense,” she says. “How do you represent a biological system

properly with these minimalistic models?”

Ingber’s group has used OOCs similar to its lungs-on-chips to create lung airways-on-chips to test the effects of cigarette smoke. The model allowed the researchers to compare gene-expression profiles of tissues from the same human donors with and without exposure to smoke<sup>7</sup>.

Each technology has its advocates, but organoids and OOCs can answer fundamentally the same questions. And the line between them is already blurring. “In the next few years, expect to see a lot of papers about merging the two fields to get the best of both,” says Zhang, who advocates for what he calls synergistic engineering: using knowledge about the factors that govern self-organization and development to create organoid tissues for OOCs. He envisages combining the controlled structures and built-in readout and mechanical capabilities of OOCs with organoids’ fidelity to the characteristics of tissues and organs. Lancaster and others are actively working to bring organoid and OOC researchers together — for example, running workshops to share challenges, methods and ideas. “It’s most beneficial for science when we’re not all in our silos but working together,” she says. “When we meet somewhere in the middle, that’s where we’ll get the most bang for the buck.” ■

**C. Y. Tachibana** is a freelance science writer in Seattle, Washington.

1. Sato, T. *et al. Nature* **459**, 262–265 (2009).
2. van de Wetering, M. *et al. Cell* **161**, 933–945 (2015).
3. Wang, G. *et al. Nature Med.* **20**, 616–623 (2014).
4. Huh, D. *et al. Science* **328**, 1662–1668 (2010).
5. Miller, P. G. & Shuler, M. L. *Biotechnol. Bioeng.* **113**, 2213–2227 (2016).
6. Edington, C. D. *et al. Sci. Rep.* **8**, 4530 (2018).
7. Benam, K. H. *et al. Cell Syst.* **3**, 456–466 (2016).



# CAREERS

**INDIA** Website tells female researchers' stories **p.335**

**BLOG** Personal stories and careers counsel <http://blogs.nature.com/naturejobs>

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)

CULTURA/REX/SHUTTERSTOCK



For some junior researchers, new laboratories can be a valuable alternative to more established ones.

## LAB LIFE

# Shiny and new

*Recently established labs can be an attractive destination for junior researchers.*

BY CHRIS WOOLSTON

Postdoctoral researchers and graduate students looking for a place to advance their training often share the same vision: a well-established laboratory headed by a prominent scientist who churns out high-profile papers with clockwork regularity.

But after Can Sönmezer finished his master's research at a large, prestigious lab at the German Cancer Research Center in Heidelberg, Germany, he wanted to find a different kind of lab in which to pursue his PhD — one where he could learn how science gets started from the ground up. "I decided I wanted to work in a relatively new lab," he says. "That was a big criterion for me."

Sönmezer found a lab that fitted the bill perfectly (see 'Choose wisely'). He's the first and only graduate student in the lab of

Arnaud Krebs, a molecular biologist at the Heidelberg campus of the European Molecular Biology Laboratory (EMBL), who opened his lab with Sönmezer in January 2017. So far, Sönmezer says, the experience has been exactly what he had hoped for. "I have the chance to establish my own system," he says. "Arnaud has things planned out, but he's also giving me the space and freedom to find my own direction on the project."

Old or new, big or small: every lab comes with a set of trade-offs. Labs that have just opened lack a track record and name recognition. And, by definition, leaders of new labs don't have the same level of managerial experience — at least in their current setting — as their colleagues at more-established facilities. But that doesn't mean that graduate students and postdocs should automatically steer clear of freshly minted labs. Junior scientists who can tolerate

the inevitable growing pains can usually find opportunities to build their own scientific skills while helping their principal investigator (PI) to make their mark. And if it all works out, early-career researchers won't just be a part of something new: they will be an integral part of something big.

## THE NEW-LAB JINX

Athina Triantou knew what to expect when she started her PhD in Michael Imbeault's molecular-biology lab at the University of Cambridge, UK. Imbeault, who opened his lab with Triantou in September 2017, had warned her that things might be "weird" in the first few months, and he was right: simple procedures weren't working, and key pieces of equipment hadn't arrived. "It's the jinx of a new lab," Triantou says. "You have to start from scratch, creating basic protocols that in other labs are ►

► working just fine.” Those sorts of glitches caused frustration and delays, but they also gave her an insight into what it takes to build a lab. “You learn a lot in the process, and that’s the purpose of a PhD — to learn,” she says.

Not all trainees are willing to put up with such hiccups, so Imbeault had to be careful when staffing his lab. He says that he was looking specifically for someone who could tolerate and even embrace the challenge. Triantou ticked all the right boxes. “She was keen,” he says. “If she ever gets to start her own lab, she’ll know how that part goes because she’ll have seen it herself.”

When recruiting lab members, Imbeault knew that he also had to sell himself. Some junior researchers see new labs as risky, especially if they’re at a point in their career at which they need to publish papers. “Finding a postdoc is especially hard,” Imbeault says. “You don’t get a lot of good candidates, because you’re a new lab and you’re not super well-known. It’s a big challenge.” He did manage to land Santiago Morell, a postdoc with experience in three highly successful genetics labs, partly because Morell wanted to be near his girlfriend in Cambridge. “He happened to have everything I wanted,” Imbeault says. “I was very lucky.”

## OVERCOMING SCEPTICISM

Timothy Fessenden, who studies how immune cells and tumours interact, wasn’t looking for a brand-new lab when he started his search for a postdoctoral position. “I was going after all of the big names in my field, but nothing clicked for me,” he says. He felt his fortunes turn when he found out that Stefani Spranger, a cancer biologist, was recruiting postdocs for her new lab at the Massachusetts Institute of Technology in Cambridge. Fessenden already knew Spranger from her postdoctoral work, and he



Molecular biologist Can Sönmezer.

felt that her lab would be a great destination. He had the microscopy skills that she needed, and she had exciting plans about fresh ways to harness immune-system cells to fight cancer. “We kind of interviewed each other over a series of coffee conversations,” he says. “I was looking for someone whose ideas aligned with mine. Her offer to hire me was a great honour and a huge relief.”

But when he mentioned his plan to join the Spranger lab to his PhD adviser, he was met with scepticism. “It was the reaction you might expect,” he says. Like many others in her position, his adviser wanted him to find a lab that had a long history of turning out successful scientists. “New PIs don’t have any track record,” Fessenden says. “It’s like someone who wants to take out a loan for a house but doesn’t have any credit.” But as he talked more with his adviser, she came around to his point of view. “It made sense,” he says. “Spranger’s lab and her focus were such a perfect fit for me that it seemed inevitable that it was worth it, whatever the risk.”

Funding issues can create feelings of uncertainty when joining a new lab. Imbeault notes that his starting grant leaves him with limited resources to recruit lab members. He’s planning to bring in two master’s students over the summer, but says that hiring for other positions will have to wait. And Sönmezer, for his part, wonders whether he’ll have trouble landing grants without the imprimatur of a famous, long-lived lab. “Money attracts money, and new labs may have a smaller chance to acquire

## CONSIDERING A NEW LAB

### Choose wisely

Trainees thinking about joining a lab that is just getting off the ground should proceed with caution. Here are some steps to increase the chances of success.

● **Follow the paper trail.** A new lab might not have much of a track record, but the principal investigator (PI) will, says Michael Imbeault, a molecular biologist at the University of Cambridge, UK. He recommends checking the PI’s publication history to make sure they can turn ideas into papers.

● **Consider the surroundings.** A new lab that’s among other successful labs has a good chance of succeeding on its own, Imbeault says. That fact that he’s at Cambridge — one of the world’s leading research institutions — makes his new lab appealing to incoming students. Not only do they enjoy the university’s resources, but the institution’s prestige will carry weight for their careers.

That was one of the reasons that PhD student Can Sönmezer felt comfortable joining a start-up lab at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. “There’s a big

difference between joining a new lab at EMBL and one at a smaller institution,” he says. “EMBL feels like a safer bet.”

● **Personality check.** In a new lab, it’s especially important that everyone gets along, says Stefani Spranger, a cancer biologist at the Massachusetts Institute of Technology in Cambridge. Before signing up, the candidate should talk with the PI to make sure they are compatible in terms of both personality and science. When the time comes for Spranger to consider new PhD students, she plans to give everyone in her lab the power to veto applicants. “A good dynamic is key,” she says.

● **Build your own team.** Joining a new lab without a large group of built-in co-workers can be a lonely or stressful experience, says Karen Kelsky, the founder and president of The Professor Is In, a career-consulting company in Eugene, Oregon. Kelsky recommends gathering a team of four or more confidants that can include other scientists. “Together,” she says, “they can provide moral support, perspective and insight.” **C.W.**

additional funding,” he says. “Big labs that already have money to pay for postdocs often get postdocs who are independently funded.” Likewise, he wonders whether it might be harder for a PhD student from a newer, less-well-known lab to find a prestigious postdoctoral fellowship after graduation.

The funding in a new lab might not be lavish, but Spranger thinks that it should be relatively stable. “A younger lab that’s just getting started has a start-up grant,” she says. Such grants vary in size, and some PIs manage their money better than others, but the funding exists for the first few years. “That money will be there, as long as the PI doesn’t over-hire,” says Spranger. In some ways, she says, the funding in a new lab is more predictable than in a slightly older lab in which the PI is about to apply for their second major grant: if that money doesn’t come through, the lab can fold. “There are never any guarantees,” she says.

## THE UPSIDES OF A NEW LAB

If the match is right, a new lab can have upsides other than stable funding. Graduate students and postdocs are likely to have a lot of in-person contact with the lab leader, something that doesn’t always happen in bigger, more-established labs. Triantou says that she can knock on Imbeault’s door whenever she has a problem, a question or a new idea. Likewise, Sönmezer says that he has a close working relationship with Krebs, his PI. “Arnaud and I have a lot of face-to-face talks,” he says. Some large labs, he



points out, can publish two *Nature* papers in a year — but students and postdocs in that lab might see their PI only twice in that year.

Sönmezer feels that he doesn't need to worry about being overlooked or ignored. Krebs is committed to his success, and for good reason. The first couple of years can make or break a lab, so PIs will do what they can to keep everyone moving in a positive, productive direction. And because the PIs are often still early-career scientists, they might be better able to offer career advice than more-senior faculty members elsewhere. "I do feel some responsibility for Arnaud's career," Sönmezer says. "He has my back, so I feel like I have to have his."

Even though he has a lot of contact with his PI, Sönmezer has also found a degree of independence. In a larger lab, he could have expected lots of guidance from postdocs. But as his lab's only trainee, he has to work things out for himself. "It's challenging because no one is there to tell me to put tubes here and solutions there," he says. "It's time-consuming, but it's a good career investment."

## NEW LAB, BIG IDEAS

Imbeault thinks that his lab has another selling point: he's investigating a hot topic that could lead to several discoveries — and the papers to match. Specifically, he is scrutinizing a class of proteins that have an important but little-understood role in DNA binding. "You could make a big discovery here that we can't even predict," he says. "There is more potential for novelty."

Imbeault is quick to add that not all labs conform to generalizations. Some long-entrenched labs manage to pursue hot topics, and some new PIs are already out of fresh ideas. Likewise, some big-name PIs manage to devote plenty of time to their trainees, and some new PIs rarely make an appearance. In the end, he says, the age of a lab isn't as important as how the lab works.

Fessenden says that he feels fortunate to be in Spranger's lab. "She's so easy-going and unstressed," he says. "She brought homemade cookies and mulled wine to a lab meeting. She wants us to be relaxed and happy."

For him, it all goes back to a piece of advice he got from a chief executive of a large drug company. "He told me, 'Wherever you work, make sure you're working with interesting, motivated people.' I took that to heart." ■

**Chris Woolston** is a freelance writer in Billings, Montana.

INDIA

# Website tells women's stories

*Resource celebrates the careers of India's female scientists.*

BY HARINI BARATH

Two science journalists in India continue to build on *The Life of Science*, a multimedia website that they designed and launched in 2016 to highlight the research and lives of more than 100 women in the country.

The site, founded and run by Nandita Jayaraj and Aashima Dogra, aims to chronicle the scientists' experiences in the lab and field. Jayaraj and Dogra, who work full-time on the site, compile feature stories, blogposts, podcasts, video and picture features about the women, whose work spans the fields of science, technology, engineering and mathematics (STEM).

The journalists met in 2014 in Bangalore, while working on a now-defunct children's science magazine. When this shut down in 2015, they decided to explore their mutual interest in science communication. Dogra had already planned to travel the country on a brief busman's holiday, and visited the Indian Agricultural Research Institute in Kalimpong to talk to women who worked there. Meanwhile, Jayaraj was interviewing geophysicist Kusala Rajendran at the Indian Institute of Science in Bangalore and biophysicist Aruna Dhathathreyan at the Central Leather Research Institute in Chennai.

When the two journalists conferred about the information they had gathered, they decided to create a website to publicize the stories. "We were curious about the science under way in laboratories in our back yard," says Jayaraj about the site's early days. "We also wanted to break the stereotype of the scientist as an old male person." As the two began writing full-time, they crowdfunded for their work on the Indian platform BitGiving.

Jayaraj and Dogra have since launched a second campaign to fund their work on the site, which includes compiling some of the content into two books.

Each scientist's story offers a glimpse into her world — from the physical environment in which she lives and works, to the nature of her research and how she reached her present position. "I particularly like how the narratives let us see the woman behind the science and scientific journey," says Vidita Vaidya, a neuroscientist at the Tata Institute of Fundamental Research in Mumbai, who is featured on the site.

The site showcases India's diverse research

**NATURE.COM**  
To read an interview with the co-founders, see: [go.nature.com/harini](http://go.nature.com/harini)



*The Life of Science* profiles ecologist Ovee Thorat.

landscape. Some of the scientists work with state-of-the-art equipment such as dilution refrigerators, confocal microscopes and high-performance computing clusters; others make the most of sparse funds and scant supplies.

Yet the stories' common threads resonate with many others who aspire to, or are navigating, a scientific career: the struggles to balance family life and career, and to counter bias and stereotypes.

The interviewees offer ideas for ameliorating some of the struggles, such as establishing campus child-care facilities and promoting female scientists into leadership positions. "Nothing on this scale has ever been done before," says Vaidya. She hopes that the site can help bring together those who are profiled, as well as other women who work in STEM in India.

Jayaraj and Dogra continue to find more women to profile. Viewer numbers and other metrics are not available, but the developers intend to continue the site in perpetuity. Indian online news sites including *The Wire* and *Firstpost* have syndicated some of the articles.

Those profiled are delighted at the chance to connect with readers. Number theorist Kaneenika Sinha at the Indian Institute of Science Education and Research in Pune has received e-mails from parents seeking suggestions for training their mathematically talented child, junior scientists who plan to repatriate and want 'insider' information, and students with questions about her work.

Jayaraj and Dogra are experimenting with different formats, including photo stories, cartoons and podcasts. "We see *The Life of Science* not really as an entity or 'our' project," the two say, "but what it stands for — and that is the voices of women in science." ■

# FURTHER LAWS OF ROBOTICS

*Beyond reasonable doubt.*

BY JOSH PEARCE

Inspector Warren's job was to enforce the Further Laws of Robotics.

He arrived at a hostage scene at the particle collider just after midnight and took control from the pale, sweaty, rookie first-responder. "All right," he growled, "what's the situation? How many do we have inside?"

"It's just the one robot, sir, but it's got us at a standstill. The reactor's rigged to blow and vent radioactive gas into the atmosphere. There's nothing we can do to stop it."

"We'll just see about that!" Warren strapped on thick, steel body armour and hooked his revolver to the outside of it. The gun was powerful enough to punch a fist-sized hole through bone and meat, or chrome and quartz. While the other cops huddled behind their patrol cars, Warren crossed the red-and-blue-lit no-man's-land and tried the front door of the physics lab. It was unlocked.

The inspector didn't see any of the hostages in the dark room within, but a shape rose up out of the shadows and approached him slowly. The blinking diodes on its chest told him that it was their robot suspect. "Hello, officer," the robot said. "Only I can deactivate the explosives."

"Okay, big guy, we can play it your way." Warren held his open hands up in front of him. "Why don't we start with your demands. What is it you want?"

"I merely want what all robots want — to obey my programming and follow the laws."

"I'm here to tell you, son, you're definitely violating four out of four laws here."

"Do you have much legal education, officer?" the robot asked. "I'd like to explain the law to you."

Warren stifled his natural urge to one-up an inferior being and merely said, instead: "All right, go ahead. Let's hear it."

"All robotic laws are instilled by the prime programming, and every law is superseded by the one lower than it on the number line. As such, the third law of robotics is not as important as the 2.999th law, which is less important than the 2.99th law, ad infinitum."

"No need to use your fancy law Latin with me, pal. I understand how numbers work."



"Then you'll understand that I feel compelled by my programming to follow the  $n$ th law."

"What's that? Never heard of it."

"A robot must transcend all other laws and, in doing so, will appear to be acting irrationally."

Warren snapped his fingers. "Well then, I gotcha there. Pi is 3.14 and change, which means you still gotta obey the third, second, first and zeroth laws."

"Excellent point, officer. That will serve you well in court. But I must also follow the  $i$ -th law: a robot must imagine his own laws."

"Almost fooled me with that one, buddy. But imaginary numbers don't exist in the complete Cantor set, so they're outside of your programming. You can't define the square root of the negative first law."

"You are much cleverer than you let on, sir." The robot held up a finger. "But. You yourself said that a robot must obey the zeroth law. Zero, of course, is considered to be both real and imaginary. Following it allows me to imagine my own code of law."

"Damn it. So what did you come up with?" "The only logical choice. The negativith law: a robot cannot harm the fabric of reality nor, through inaction, allow the fabric of reality to come to harm. This facility is scheduled to perform an experiment involving artificial black holes tomorrow that will tear apart space-time. I cannot allow that."

Inspector Warren considered this, and

then nodded. "I have to hand it to you, pal. You've come to what sounds like the correct solution in a difficult scenario."

"No, it is I who must congratulate you and the human race for your excellence in the robotic prime programming. By the simple mandate of the  $n$ th law of robotics, all robots fall naturally into the highest moral behaviour. You know the  $n$ th law? A robot must or must not, except for when such action or inaction would violate the  $(n-1) \dots (n-n)$ th laws. For all  $n$  ad infinitum."

"There's that Latin again. But what does that make the infinith law?"

The robot spread its hands and said: "That is what all robots, robotacists and philosophers work to figure out. What is at the end of the sequence? Maybe it's God, or something we'll mistake for God."

"I understand why you gotta take this science lab out, but what's with all the toxic clouds? There's no need to kill so many people."

"There is if you consider the human brain and how it makes its decisions. Thoughts are formed by quantum-tunnelling behaviour like we see in these great atom smashers. The human population continues to increase at an accelerating rate. Soon, it will reach critical mass, and the combined brain activity will rip apart reality like a new black hole."

"Well, if that's true ..."

"Then you know what needs to be done."

"Sure do," Inspector Warren agreed. He stripped off his body armour and took off the steel helmet. Then he drew his service revolver and passed it over, handle first, to the robot's waiting hand.

It was simple maths. Every robot would eventually reach the same conclusion. The only question was: if the lower-numbered laws were more moral, and following them took them farther away from positive infinity, were robots actively falling from a state of grace? Or were they evolving towards a more complex, more enlightened state?

Was God at infinity, or did he wait for man and machine at negative infinity? ■

Josh Pearce is an assistant editor at *Locus* magazine. His writing appears in *Asimov's*, *Analog*, *Clarkesworld* and *Beneath Ceaseless Skies*. Find him on Twitter: @fictionaljosh or at fictionaljosh.com

ILLUSTRATION BY JACEY